AD_____

Grant Number DAMD17-94-J-4509

TITLE: Massachusetts Cancer Control Evaluation Project

PRINCIPAL INVESTIGATOR: Susan T. Gershman, M.P.H., Ph.D.

CONTRACTING ORGANIZATION: Massachusetts Health Research
                           Institute, Incorporated
                           Boston, Massachusetts   02108

REPORT DATE: October 1997

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
               Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
                         distribution unlimited

The views, opinions and/or findings contained in this report are
those of the author(s) and should not be construed as an official
Department of the Army position, policy or decision unless so
designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE October 1997	3. REPORT TYPE AND DATES COVERED Final (23 Sep 94 - 22 Sep 97)

4. TITLE AND SUBTITLE
Massachusetts Cancer Control Evaluation Project

5. FUNDING NUMBERS
DAMD17-94-J-4509

6. AUTHOR(S)
Susan T. Gershman, M.P.H., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Massachusetts Health Research
  Institute, Incorporated
Boston, Massachusetts 02108

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200* Currently, there are no proven ways to prevent breast cancer. However, many studies have shown that mammography screening programs can reduce breast cancer mortality 30 to 40%. It is imperative, therefore, to create reliable systems to monitor the effectiveness of screening programs. Because accurate measures of screening effectiveness are not available, we analyzed the proportion of cases with advanced disease at diagnosis within each geographic area as a proxy measure of screening efficacy. A cluster of tracts with excess proportions of late stage diagnoses is evidence of poor screening within that geographic area. As part of the Mass. Breast Cancer Control Evaluation Project, these proxy measures of screening have been incorporated into a Geographic Information System (GIS) along with other relevant social and economic data from the 1990 Census. Spatial scan statistical analysis determines whether significant excesses of late stage cancers are clustered geographically. When significant clusters are found, GIS is used to create thematic maps of the racial/ethnic, educational and economic characteristics of those areas, along with the location of mammography sites. Such information can help in designing mammography screening programs tailored to the unique characteristics of those regions where screening is currently ineffective.

14. SUBJECT TERMS Breast Cancer
Cancer control evaluation, geographic information systems, geocoding, spatial scan statistics

15. NUMBER OF PAGES
80

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

<u>FOREWORD</u>

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

_____ Where copyrighted material is quoted, permission has been obtained to use such material.

_____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

_____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

_____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

_____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

_____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

_____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

_____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.
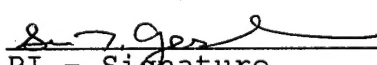
PI - Signature   10/21/97
                                  Date

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION:

Purpose

This study describes a system to assess the efficacy of breast cancer screening. Since direct measures of screening are not available, this project uses proxy measures based on diagnostic staging. The information provided by this system can be used by public health officials not only to identify geographic regions where screening is inadequate, but also to identify and characterize the educational, economic, and racial/ethnic background of persons residing in these regions and to tailor interventions to fit the characteristics of the local population. The system for conducting such an assessment can also provide concomitant information about the location of mammography units, display data geographically on maps, and allow for querying of the displayed data so as to obtain information on any location.

Background

Breast cancer is the leading cancer among Massachusetts females, accounting for 31.1% of all newly diagnosed cancer cases between 1982 and 1994. During this time period, there was an alarming rate of increase in diagnosed cases, prompting state government officials in May 1992 to declare the disease an epidemic. Between 1982 and 1991, the age-adjusted incidence rate increased nearly 35%, from 90.0 cases per 100,000 females to 121.7 cases per 100,000 females. Since 1991, rates have declined slightly, to 115.3 per 100,000. Nationally, rates increased 23% between 1982 and 1994, from 89.3 per 100,000 to 109.7 per 100,000. The incidence of breast cancer in Massachusetts is about 6% higher than rates from the SEER program.

Since there is no effective primary prevention strategy for breast cancer, secondary prevention, through mammography screening and early detection, remains the only way of controlling breast cancer and improving survival. Screening has been shown to reduce breast cancer mortality 30 to 40% among women aged 50 and older (Collette, 1992; Shapiro, 1982; Habbema, 1986; Chu, 1988). A large scale randomized controlled trial in Sweden reported a 30% reduction in breast cancer mortality for women aged 40 or older attributable to mammography (Tabar, 1985 and 1992).

The use of changes in mortality rates to assess the effects of mammography screening requires costly and time-consuming follow-up of cases, however. It has been noted that breast cancer case fatality rates are about 3 to 4% per year for the first 8 years after diagnosis, and 1 to 2% per year for the next 8 years. Because of this, "trends in mortality would be expected to lag behind any increase in underlying risk or in the use of screening by at least 5 to 10 years" (Chevarley, 1997). This need for long-term follow-up emphasizes the need for surrogate measures, such as utilization of screening mammography and stage at diagnosis.

Researchers from the University of Massachusetts Medical Center conducted an assessment of the effectiveness of a multicomponent intervention in two communities to increase utilization of breast cancer screening by women over 50 years of age (Zapka, 1993). They found dramatic improvement in both the intervention and control groups, and concluded that participation in screening was a rapidly rising secular trend.

Furthermore, the technology of mammography screening has also improved over time. In Vermont, 34% of the cases diagnosed between 1975 and 1984 were less than two centimeters; this proportion had increased to 50% of the cases diagnosed in 1989 and 1990 (Foster, 1995). In Massachusetts, the proportion of breast cancer cases diagnosed at a local stage increased from 49.4% in 1982 to 54.4% in 1992. An additional 14.9% of 1992 cases were diagnosed *in situ*, for a total of 69.3% of breast cancers diagnosed at an early stage that year (Gershman, 1997).

A 1995 Swedish study looked at non-attenders of screening programs (Lidbrink, 1995). Non-attenders were subdivided into two groups: those who avoided mammography, and those who were screened outside the program. Of those who avoided mammography, 33% said they would never have one (definite non-attenders) and 29.5% had missed the appointment (possible future attenders), while 32% were screened outside the trial. The stage at diagnosis in the non-attenders was compared to the stage of clinically detected breast cancer in a non-screened control population. The non-attenders who did not screen elsewhere had later stage diagnosis and significantly higher mortality rates than the nonscreened control population, while the non-attenders who screened outside the trial had stages similar to the attenders. This study further confirms the premise that higher stage at diagnosis reflects poor screening.

Mandelblatt et al. (1995) found a relationship between sociodemographic factors and stage at diagnosis of breast cancer. In their work, they assessed the effects of individual-level demographic characteristics (such as age and race) and measures of social context (such as the SES of the area of residence, change in area SES, and access to mammography) on breast cancer stage at diagnosis among New York City residents. As in this study, they had no individual measures of SES, but used census-tract data as surrogates. They found that African-American women were 25% more likely than White women to have late-stage breast cancer, that area mammography capacity was independently associated with stage of diagnosis, and that area SES was independently associated with late-stage disease.

The relationship between low SES and poor health has been well documented, yet difficult to explain after taking into account important confounders such as health habits and access to health care. Racial differences are also documented. As many African-Americans and Hispanics are economically disadvantaged, researchers have tried to explain the racial differences by the socioeconomic differences, but the interrelationship between race and SES may be too difficult to unravel with traditional adjustments for current income and education. Tools which assess wealth in addition to income may help, but it seems that other difficult-to-measure factors, such as chronic stressors which

reflect the lifetime impact of discrimination, may be associated with the excess of disease in racial minorities. (Guralnik, 1997)

This project is directed at evaluating screening efficacy across the entire state of Massachusetts, and builds upon the surveillance systems of Kerner (1984) and Andrews (1994). Kerner and his colleagues examined geographic variation in disease incidence and mortality in relation to census variables in an attempt to target screening programs, while Andrews and his colleagues combined mortality and census data to target cancer screening programs on a geographic basis. This project's system integrates data from a cancer registry with data from the census, along with other health information such as location of mammography screening sites, into a single geographical information system (GIS). Dangermond (1990) defined a geographical information system as "an organized collection of computer hardware, software, geographic data and personnel to efficiently capture, store, update, manipulate, analyze and display all forms of geographically referenced information". Maguire (1991) argues that "it is the ability to organize and integrate apparently disparate data sets together by geography which make GIS so powerful. The spatial searching and overlay operations are a key functional feature of GIS." Some elements of the GIS used in this study are diagrammed in Figure 1 and described below.

Previous Work

Year 1 activities focused on examining the distribution of breast cancer in Massachusetts and throughout the US. Using data from Massachusetts, Connecticut, California and the National Cancer Institute's Surveillance, Epidemiology and End Results program, trends in cancer incidence, staging, mortality and mammography screening were explored, and integration of these data sources begun. Project staff also analyzed census data, prepared population data for multiple geographic units of analysis and time periods, and examined correlations between various socioeconomic factors. Additionally, a master file of data sources was compiled in preparation for developmental modeling, and began the statistical modeling.

Year 2 activities focused upon completion of the statistical model. In addition to the socioeconomic variables created from measurement modeling of the census tract measures, other known covariates were analyzed in Year 2. Spatial scan statistical techniques were utilized to examine the distribution of stage at diagnosis of breast cancer cases, and to identify areas of the state with clusters of late-stage diagnoses. Data sets analyzed were incorporated into a mapping software package, allowing the user to view geographic representations of data distributions. Through this technique, an area of the state was identified in which there appeared to be an excess of late-stage diagnoses. Concomitant socioeconomic data was then made available so the community could be characterized and areas of intervention identified.

BODY:

Methods

Since January 1, 1982 all new cases of cancer diagnosed in Massachusetts residents have been reported to the Massachusetts Cancer Registry (MCR), a Division of the Massachusetts Department of Public Health. Each report to the registry is recorded on a standardized form to obtain comparable information from case to case about the type, histology and stage of the disease. Forms also include demographic information, including the patient's age, race, occupation, smoking status, and address at the time of the diagnosis. For this study, breast cancer cases diagnosed between 1982 and 1992 were aggregated by census tract and integrated with geographical information, such as the location of 1177 census tracts, the location of 351 minor civil divisions (MCDs)[1], the location of 296 mammography machines at 218 sites, and the boundary files for each of 27 Community Health Network Areas (CHNAs)[2]. Breast cancer data were aggregated into two five-year periods, 1982-1986 and 1987-1992. While data from the first period was used to demonstrate the system and to identify areas of high or low screening efficacy, substantive findings and the consistency of those findings over time can be cross-validated with data from the second period. As diagrammed in Figure 1, data were also extracted from the 1990 Census so that tracts could be characterized according to a variety of social, economic, and demographic indicators, such as educational attainment, race/ethnicity, per capita income, employment levels, and the distribution of occupational categories.

Measures

Accurate measures of mammography screening are not generally available. A Wisconsin study compared levels of mammography screening using data from the Behavioral Risk Factor Surveillance System (BRFSS) to data from records of mammography sites. The two data sources showed similar trends, but large and consistent discrepancies in terms of the actual number of mammograms performed (Lantz, 1995). Estimates of screening from BRFSS data consistently overestimated rates of screening by about 20% as compared to data obtained from the mammography sites.

Since direct measures of screening are generally not available, this project uses proxy measures. One proxy measure suggested by Roffers and Austin (1993) is based

---

[1] MCDs are equivalent to the 351 incorporated cities and towns in Massachusetts. Although the data in this paper have been aggregated at the level of the census tract, it is possible to disaggregate further to the block group level, or aggregate to the MCD level.

[2] CHNAs are a Massachusetts Department of Public Health designation for aggregations of cities and towns. CHNAs are used to develop health networks consisting of consortia of health care providers, human service agencies, schools, churches, advocacy groups and members of the public of all ages. These networks identify and assess health needs in their communities, and evaluate responses to these needs. The major foci of the networks are increasing access to care, efficiency of health services, and communication and collaboration among health care and human services providers in the area.

upon the proportion of cases diagnosed at an *in situ* or localized stage. Boss and Suarez (1990) also suggested using the ratio of *in situ* diagnoses to all invasive cases as a measure to evaluate screening programs. Roffers and Austin maintain that if at least 5% and up to 15 or 20% of newly diagnosed cases are *in situ* for a community, mammography screening can be judged as satisfactory. The MCR began collecting data on cases diagnosed at the *in situ* stage in 1992 (previously, only invasive cancers were required to be reported). Stages used in this analysis are: Stage 1 (localized disease), Stage 2 (regional spread of disease), or Stage 3 (remote or metastatic disease).

Figure 1. Elements of a Geographic Information System (GIS) for breast cancer control evaluation.

Geographic Data	Cancer Registry Data
27 CHNAs	Census Data
351 Towns	Mammography Data
1177 Census Tracts	

Cancer Registry Data	1990 Census Data	Mammography Data
Incidence	Race/Ethnicity	# of Sites
Stage at Diagnosis	Education	Location
	Economics	
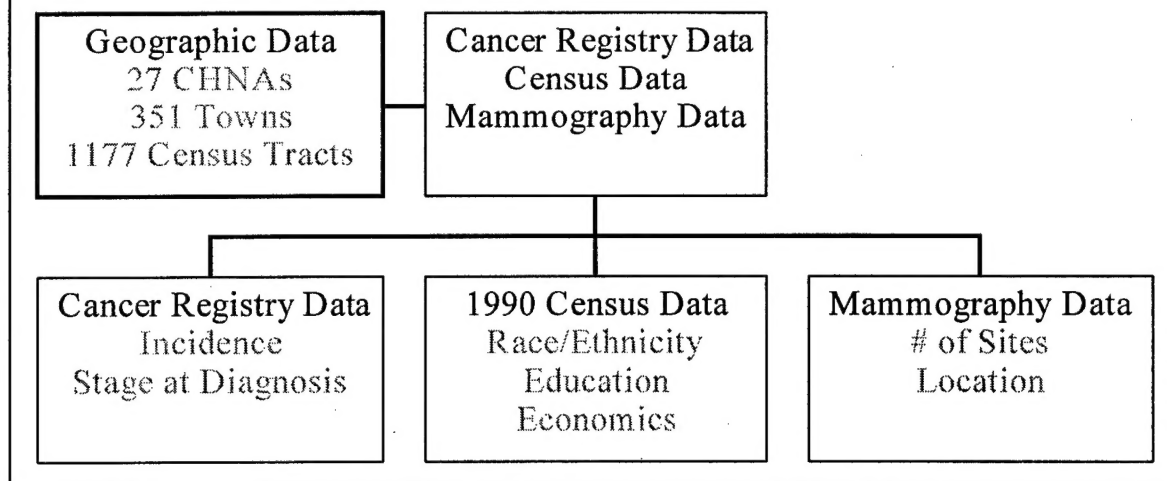
Cases were first aggregated at the census tract level because of concerns that use of higher levels, such as towns or CHNAs, would mask the broad variation found within towns and within CHNAs. The proportion of female breast cancer cases in each census tract in Massachusetts diagnosed at Stage 1, Stage 2, or Stage 3 was computed. Assuming that earlier diagnosis reflects better screening, tracts with higher proportions of Stage 1 cases (localized disease at diagnosis) were seen as having better screening than tracts with lower proportions of Stage 1 cases. Conversely, census tracts with higher proportions of Stage 3 cases (remote or metastatic disease at diagnosis) were seen as having poorer levels of screening than tracts with lower proportions of Stage 3 cases.

Statistical Analysis

A variety of univariate statistical methods were used to describe the occurrence of cancer within a region or across the state of Massachusetts, and to describe social, economic and demographic variables. Bivariate relationships were analyzed using chi-

10

square and Pearson correlations; polychoric and polyserial correlations were also used to study associations, but these analyses are not reported here. The relationships between cancer data and sets of social, economic and demographic variables were examined in a variety of ways, including multiple regression analysis and discriminant function analysis. The spatial scan statistics technique (Kulldorff, 1995) was used to test whether certain geographical regions contained clusters or excess numbers of Stage 1 or Stage 3 cases. Spatial scan statistics determine whether the higher numbers of Stage 1 or Stage 3 cases occurring in some regions exceed the number of cases attributable to chance variation. Regions with statistically significant excesses of Stage 1 cases could be viewed as screening more effectively, and regions with statistically significant excesses of Stage 3 cases could be seen as deficient in their screening programs.

Preparation of Population and Demographic Data

One of the research objectives of this project has been to determine whether or not statistically significantly relationships exist between the incidence of breast cancers among Massachusetts women and sociodemographic characteristics of the geographic areas of their residence. Several population data sets were created for use in these analyses.

There are two types of population information available: 1) population counts and estimates, i.e., the numbers of persons by gender, age, and race/ethnicity residing in specified geographic areas. Counts are reported decennially by the Census Bureau for 100 percent (ideally) of the population; estimates are made for inter- and post-censal years; and 2) demographic information, i.e., social and economic characteristics reflective of lifestyles and living conditions. This information was collected in 1990 from a 16% (approximately) sample of the population and then generalized to the whole.

The elemental geographic units are also of two types: 1) Census Tracts (CTs) and 2) Minor Civil Divisions (MCDs). Separate files were created for the CTs and the MCDs, because tracts are not consistently part of towns. In metropolitan areas, CTs are generally identified with a given MCD, and they can be aggregated to specific neighborhoods or other localities within the MCD. However, in more sparsely settled areas (e.g., western Massachusetts), tracts may be shared by two or more towns.

Census tracts are relatively permanent subdivisions of a county. They are identified by a basic 4-digit number and may have a 2-digit suffix, in which cases they are referred to as 6-digit tracts. Statewide, there are 1,331 census tracts, including 4 block numbering areas (BNAs) in Nantucket, which can be treated as CTs for our purposes. Tract numbers range from 0001 to 9351, and are unique within a county. Census tracts were delineated by the Census Bureau in Barnstable, Dukes, and Franklin counties for the first time in 1990; however, some tract numbers already used in Suffolk County (viz., Boston) were inadvertently assigned to tracts in Barnstable and Franklin counties. In these cases, the county code is a necessary prefix to identify the unique census tract.

11

Minor civil divisions are the primary political or administrative divisions of a county. In Massachusetts, they are the 351 cities/towns which have locally elected governments.

Population counts for 1980 and 1990 were provided by the US Census Bureau. They were categorized by gender and five-year age groups for each Massachusetts city/town. These were the basis for the linear interpolation of annual gender/age values per MCD for the years 1981 through 1989. Subsequently, estimates of the 1995 populations by gender/age per MCD were provided by the Massachusetts Institute for Social and Economic Research (MISER) at the University of Massachusetts, Amherst. These were used with the 1990 census counts to derive comparable annual interpolations per MCD for 1991-1994. Each record is coded to permit geographic aggregations of data for Community Health Network Areas (CHNAs), counties, and other collections of MCDs. These 1980-1995 gender/age populations provide the necessary denominators for various age-adjusted rate determinations for cancer incidence and mortality over time among geographic entities.

Census counts of gender/age distributions among Massachusetts census tracts for 1980 and 1990 were also obtained. No inter-censal interpolations were attempted, however, because of small tract populations and resultant concerns about validity and reliability of such estimates. Furthermore, there are no estimates of 1995 tract populations, which are required for any post-censal interpolations for the years 1991-1994.

For the initial tests for statistical association of cancer incidence and demographic factors, age-adjusted rates of breast cancers were calculated for MCDs. It soon became clear that census tracts could provide a more suitably refined geographic unit of analysis than could whole cities/towns. Socioeconomic and demographic items from the 1990 census were selected on the basis of demonstrated statistical association with different lifestyle characteristics. It was felt that sociodemographic information could be of particular value in characterizing geographic areas of interest to health planners. Specific target populations could be more readily located geographically; possible problems in communicating effectively might be anticipated and appropriate program modifications considered.

A file of breast cancer incidence data for 1982-1986, including stage at diagnosis, was created. Files of sociodemographic characteristics for MCDs and for CTs were each matched with the incidence file to produce two preliminary work files: TWLFSTYL (town-level data; n=351) and TRLFSTYL (census tract-level data; n=1177). The first file contained a record for each MCD, including the few in which no breast cancers had been diagnosed during the 1982-1986 period. The second file consisted of only tracts in which a breast cancer case had been reported. Of the 1331 census tracts in Massachusetts, including Nantucket's 4 BNAs, 1177 had records.

12

Statistical analyses of the data in the TRLFSTYL file did not reveal any real relationship between the independent sociodemographic variables and the stage of the cancer at diagnosis. There were, however, some interesting geographic patterns of tract clusters. The experience in dealing with the 1982-1986 information has been invaluable as preparation for analysis of the 1987-1994 data. The inclusion of *in situ* cases, begun in 1992, is expected to enhance the value of stage classification as a measure of the effectiveness of breast cancer detection.

Further details of file development and file formats are provided in Appendix A.

Spatial Scan Statistics

The spatial scan statistic (Kulldorff, 1995) is used to test whether certain geographic regions contain clusters of tracts with excess proportions of Stage 3 cases. This process "uses a circle of variable size and location to scan the whole map for areas with high or low rates. Using maximum likelihood estimation, a most likely cluster is chosen. The statistical significance of this cluster is then tested taking the multiple comparison into account that resulted from looking for clusters in many different locations and for many different sizes. The method will also detect and evaluate secondary clusters" (Kulldorff, 1997a). The spatial scan statistic determines whether the higher proportion of Stage 3 cases occurring in some regions exceeds the number of tracts with high proportions of Stage 3 cases that could be attributed to chance variation. Regions with statistically significant excesses of Stage 1 cases could be viewed as screening more effectively, and regions with statistically significant excesses of Stage 3 cases could be seen as deficient in their screening programs. Having identified such clusters, the GIS can be used to show on a map the location of those clusters and the social, economic and demographic characteristics of the people who live in those geographical areas.

A recent study by Kulldorff et al. utilized the spatial scan statistic to investigate clusters of breast cancer mortality in the northeastern United States. This technique searches for clusters of cases without specifying their size or location ahead of time. Unlike the traditional statistics test used to analyze clusters, the spatial scan statistics does not assess whether the number of cases is greater than would be expected in an area. If an "area is chosen because it has more cases, then this approach introduces preselection bias since the same cases are used to define the hypothesis as to test it" (Kulldorff, 1997b). This test requires that the geographic position of each county be specified by the latitude and longitude of its centroid. The spatial scan statistic tests the null hypothesis that within any age group the risk of breast cancer mortality is the same in all counties, assuming breast cancer mortality to be Poisson distributed.

Controlling for age, race, urbanicity and parity, the authors identified a 7.4% statistically significant excess of breast cancer mortality in the New York City/ Philadelphia metropolitan area, including Long Island, compared to the rest of the northeast. The authors were unable to control for age at menarche, age at menopause, age

at first birth, history of breast feeding, country of birth, family history of breast cancer, alcohol consumption, access to health care or any environmental factors. These risk factors may explain the detected clusters. Despite these limitations, the spatial scan statistic was shown to be a useful tool in evaluating distributions of cases.

Software Development

Project staff determined that it would be most efficient to modify existing applications that had already been created within the Massachusetts Cancer Registry - Cancer Information Management System (MCR-CIMS). Appendix B contains sample software interfaces from the Ad Hoc application, as well as a draft of the users' manual. This application will provide the cancer incidence files that will be imported into Maptitude (mapping software package). These files will provide information (such as incidence rates, staging distributions, age distributions, racial/ethnic classifications, and year of diagnosis) which will be displayed for the selected geographic areas.

Functional specifications for data files:

In order to use the model developed, a number of files that provide cancer incidence data, population data, census tract and city/town data need to be imported into mapping software such as Maptitude. These data files are as follows:

File A:	Population data -- four-digit census tracts and their respective 1980 and 1990 populations for 18 five-year age groups, by sex (see Appendix A)
File B1:	City/Town data (351 cities/towns) -- 1990 Census data by MCD, including race/ethnicity, education and economic variables (see Appendix A)
File B2:	Census tract data (1177 tracts) -- 1990 Census data by census tract, including race/ethnicity, education and economic variables (see Appendix A)
File C:	Massachusetts Cancer Registry data -- breast cancer incidence for 1982-1992 by census tract, age and stage (see Appendix B)

Geocoding and Cancer Registry File Preparation

Files were downloaded from the MCR's mainframe database and sent to a geocoding company. These files contained a linking field and the address fields from each record for diagnosis years 1982-1994, although the 1993 and 1994 data were not yet completely collected at that time. The company processed the files, automatically matching the addresses with earth-based geographic coordinates, and returned the files with census tract, block group, approximate latitude and longitude, zip code, MCD and various codes indicating the accuracy of the match attached (for those addresses which could be automatically indexed to known coordinates). For addresses which could not be automatically matched down to the census tract level, the company returned partial information, such as the latitude and longitude of the centroid of a zip code area or MCD.

The large, compressed files returned from the geocoding company had to be split into much smaller files so that they could be used in the PC environment. The returned data were imported into geographic information systems (MapInfo and Maptitude) and a quality assurance audit was conducted (Appendix C) on a selection of data from both small and large towns. Problems identified during this audit required correction, and resulted in significant delays.

Once corrected, the geocoded files had to be linked back to the MCR's mainframe records so that all of the breast cancer cases could be identified, edited and exported for analysis. The effect of the linking procedure would be to attach reliable geographic data (from the geocoded files) to the complete breast cancer records residing on the mainframe.

Next, the small files processed on PCs had to be converted to the .txt file format necessary for the mainframe environment. The small .txt files were then uploaded to the mainframe and appended together to produce one large file to expedite the linkage procedure. Once this large file was constructed, the linkage was performed. Those records with partial or unreliable geographic data (census tract, block group, latitude, longitude) were identified (visually and by using certain codes provided by the geocoding company). These spurious data were eliminated from the file so that only reasonably reliable geographic data would be included for analysis. The edited file was then processed so that the staging information in each record could be simplified into the format required for the analysis.

Finally, the fields (demographic information, geographic information and staging) to be used in the analysis were selected from each record, written out to a new file, downloaded, and sent with a file description to the project evaluator for analysis. Because of the quality assurance problems identified, and the necessary corrections required, this preparation of the files ultimately took from November 1996 through September 1997. As a result, computer analysis of 1987-1992 cases is just beginning.

Results and Discussion

During the years 1982 to 1986, the study period, there were 18,354 cases of breast cancer reported to the Massachusetts Cancer Registry. Of these, 9899 (53.9%) were Stage 1, 5853 (31.9%) were Stage 2, 1400 (7.6%) were Stage 3, and 1202 (6.5%) were of unknown stage. There appeared to be a relationship between age at diagnosis and stage, with 6.4% of the cases diagnosed at Stage 3 for women under 50 years of age, 7.4% for women between 50 and 64, 8.8% for women between 65 and 75, and 9.1% for women older than 75 years.

While all 18,354 cases could be geocoded for their town or 5 digit zip code location, only 15,473 could be located and geocoded to census tract location.

Nevertheless, because there are more census tracts (n=1177) than 5 digit zip codes (n=795) or towns (n=351), analysis was begun at the level of the census tract. Overall, 15.7% of the cases could not be geocoded to a census tract, and were therefore unavailable for analysis at this level. There appeared to be little variation in the percent missing by stage (all close to 15.7%). Likewise, the percent missing was consistent from 1982 to 1986. In the highest age category, greater than 75 years old, the percent missing was 19.7%, with the other age groups slightly lower than the mean. Geographically, however, there is considerable variation in the range of missing cases, with counties showing a range from 10.4% to 36.2% missing.

Figure 2 shows the distribution of the proportion of Stage 1 cases for each of the 1114 tracts reporting at least one case between 1982 and 1986. Sixty-three of the 1177 census tracts had no cases during that period. In fifty tracts, none of the cases were Stage 1, while in another 50 tracts, 100% of the cases were Stage 1. The remainder of tracts are distributed around the mean of .53, though the distribution shows a definite skewing to the right, suggesting that in most tracts a preponderance of cases are diagnosed at Stage 1. While Figure 2 illustrates the great variability among tracts with respect to the proportion of cases diagnosed as Stage 1, it conveys no information about geographic variability. Are some regions consistently higher, or lower, with respect to the proportion of Stage 1 cases diagnosed within those tracts?

In order to view the data geographically and to highlight the top 25% and bottom 25% of tracts, tracts were grouped into three categories: 1) the lowest quartile (tracts where the proportion of Stage 1 cases was less than or equal to 0.43); 2) the highest quartile (tracts where the proportion of Stage 1 cases was greater than 0.65); and 3) the middle 50% (tracts where the proportion of cases diagnosed at Stage 1 was greater than 43% and less than 65%). Figure 3 displays tracts from the lowest and highest quartiles. Even though it appears that there are clusters of dark tracts, which might suggest that those tracts are doing a poorer job of screening, especially in comparison to what appear to be clusters of light tracts, it would be a mistake to draw such conclusions because it is known that there may be relatively few cases in some tracts and therefore characterization of any given tract, even any given cluster of tracts, requires that such instability be taken into account. Furthermore, without a formal statistical test, it is not possible to conclude that any apparent clusters aren't due to normal and expected variability.

Spatial statistics were used to adjust and account for the variability and instability introduced by tracts with small numbers of cases. Kulldorff's spatial scan statistic was applied to determine whether there are clusters of tracts with excess numbers of Stage 1 cases. That is, as the scatter of high stage 1 census tracts is examined, are there clusters of tracts or regions with excessive numbers of Stage 1 cases above the numbers that might be expected due to normal statistical variation? For each tract the actual number of Stage 1 cases for that tract and neighboring tracts was compared to what might be expected given the population of Stage 1 cases for the entire state. The definition of neighboring tracts is continually enlarged in multiple statistical trials to include up to 10% of the total population of Stage 1 cases. The spatial scan statistic revealed no

Figure 2. Distribution: proportion of Stage 1 cases for 1165 census tracts.

Proportion of cases that are Stage 1

Std. Dev = .21
Mean = .53
N = 1114.00
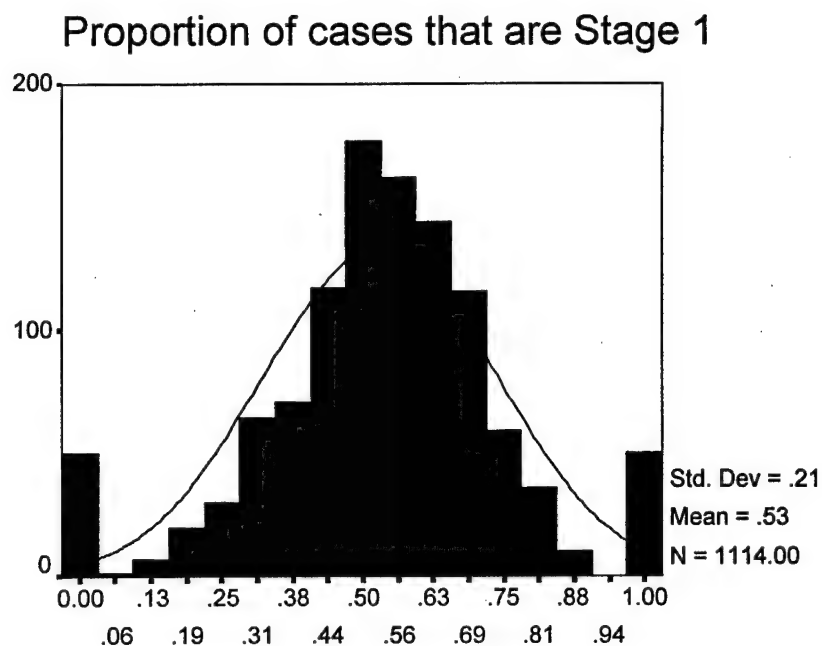
PS1826

Figure 3. Tracts in the highest quartile for proportion of Stage 1 cancers (light) and tracts in the lowest quartile for proportion of Stage 1 cancers (dark).

statistically significant clusters of tracts with excess numbers of Stage 1 cases. Thus, looking again at Figure 3, if there appear to be clusters of light tracts (tracts with a high proportion of Stage 1 diagnoses), those clusters are only apparent and can be attributed to normal variation; if they can be attributed to chance variation, it would be a mistake to attribute them to something else, such as excellence in screening.

Another question suggested by Figure 3 deals with the relationship between educational, racial/ethnic, and economic indicators available for each tract from the 1990 census and the proportion of Stage 1 cases. While the census measures correlate quite highly among themselves, there are no statistically significant correlations between any of the census measures and the proportion of Stage 1 cases.

Another way of examining the nature of the relationship between the sociodemographic variables and the proportion of Stage 1 cases was to compare those tracts in the highest quartile to those in the lowest quartile with respect to the proportion of Stage 1 cases. Table 1 shows the mean for each of the census variables for the low and high Stage 1 categories. None of these differences are striking, and none are statistically significant.

Table 1. Low and high Stage 1 tracts by selected sociodemographic variables from the 1990 Census, Massachusetts.						
Proportion Stage 1	< 9 yrs ed	College Grads	Black	Hispanic	Unemployed	Per capita Income
Lowest 25%	9.84%	26.84%	6.18%	5.39%	7.80%	$ 17,015
Highest 25%	10.57%	25.50%	4.49%	6.02%	7.64%	$ 16,272

While an examination of census tracts according to the proportion of Stage 1 cases reported provides some insights into whether tracts are doing well or poorly with respect to detecting cases early, an examination of tracts according to the proportion of Stage 3 cases may reveal whether tracts are doing well or poorly with respect to detecting cases that have already metastasized at the time of diagnosis. Figure 4 shows the distribution of tracts according to the proportion of Stage 3 cases diagnosed in residents of each tract. While, fortunately, the distribution is quite skewed, with most tracts showing a low proportion of Stage 3 cases, nonetheless there is variability, and a long tail with some tracts showing a relatively high proportion of Stage 3 cases.

As with the Proportion of Stage 1 cases, tracts were divided into three categories: 1) the lowest 42.7% (tracts where no cases were diagnosed at Stage 3); 2) the highest quartile (tracts in which 12% or more of cases were diagnosed at Stage 3); and 3) the remaining tracts, in which at least one case but fewer than 12% were diagnosed at Stage

Figure 4. Distribution: proportion of Stage 3 cases for 1114 census tracts.

Std. Dev = .12
Mean = .08
N = 1114.00
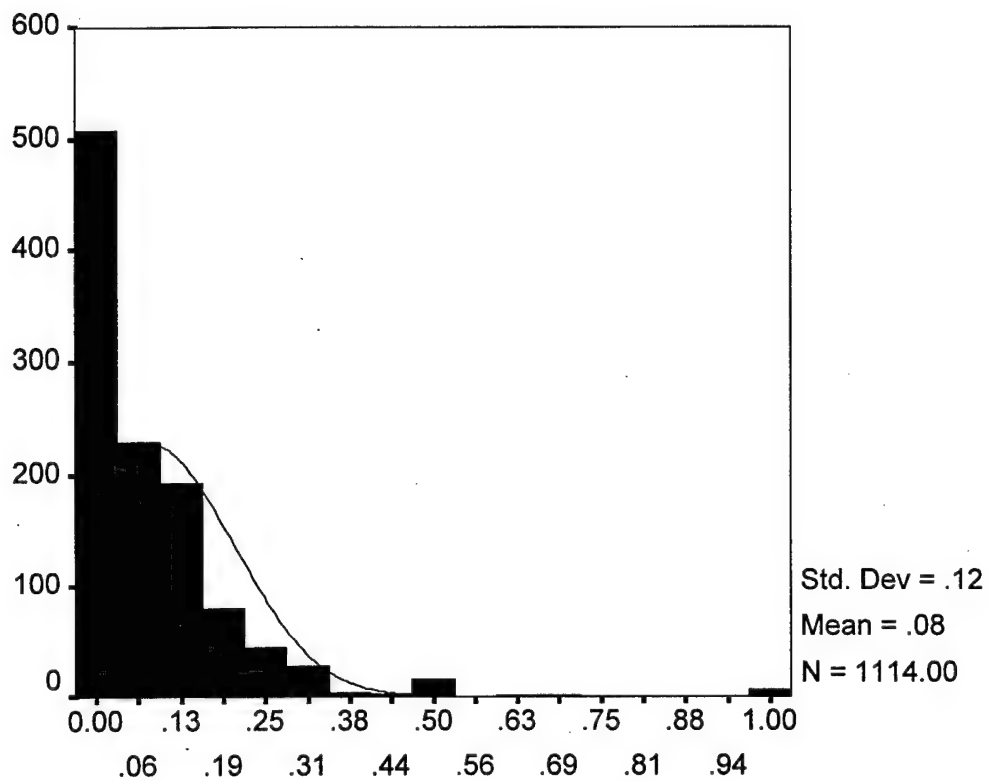
PS3826

Figure 5. Tracts with 0% (light) or greater than 12% (dark) of their cases diagnosed at Stage 3.

3. Figure 5 presents the same data geographically. While Figure 4 shows that most cases are *not* diagnosed at Stage 3, it provides no information about the location of tracts where the proportion of Stage 3 diagnoses is relatively high, and no information on whether there are clusters of tracts with higher proportions of Stage 3 diagnoses.

In Figure 5, the light tracts (42.7% of tracts) report none of their cases as Stage 3. The dark tracts are in the highest quartile, and have 12% or more of their cases diagnosed at Stage 3. Are there regions of the state where there are excessive numbers of stage 3 cases relative to the total number of stage 3 cases? Kulldorff's spatial scan statistic was applied by taking the number of stage 3 cases in each tract and comparing it to the number of cases expected if only chance variation were operating. That is, it compares the actual number of Stage 3 cases occurring in groups of tracts relative to the proportion of Stage 3 cases expected across similar groups of tracts. In this instance, there were a total of 1191 Stage 3 cases out of a total of 15,473 cases for the five year period. The spatial scan statistical analysis identified no significant clusters. The analysis would support the conclusion that any variation in the proportion of Stage 3 cases among census tracts can be explained by chance.

Because the spatial analysis by census tracts was performed excluding the 15.7% of the cases that could not be geocoded to census tracts, a spatial scan was next performed of the data from the same 1982-1986 time period, organized by town. Because all 18,354 cases were identified with one of the 351 cities or towns in Massachusetts, there were no cases missing in these analyses. There were 1400 Stage 3 cases reported. The spatial scan analysis revealed a statistically significant cluster of towns with higher than expected proportions of Stage 3 diagnoses (p=0.015). Within this cluster of towns there were 1211 total cases, of which 135 were Stage 3. Statistically, according to the spatial scan analysis, 92.37 cases would be expected, yielding a relative risk for these towns of 1.46. Figure 6 is a map with the cluster of towns highlighted. The exact same cluster was identified using the latest version of SaTScan (version 1.0.2), which also allows scanning by time and space simultaneously. To perform a space by time analysis, the data were aggregated by each of the five study years and by towns, with five years worth of data for each town.

Additionally, since all 18,354 cases could be located within a 5 digit zip code area, a spatial analysis on cases aggregated by zip code was also performed. In Massachusetts there are 795 five digit zip code areas, of which 528 reported cases. The spatial scan analysis identified a significant cluster of zip code areas with higher than expected proportions of Stage 3 cases (p=0.010). In the cluster area, shown in Figure 7, there were 1304 cases; of these, 146 were Stage 3. Under the null hypothesis, 99.47 cases would be expected in that area if chance alone were operating. The higher than expected number of Stage 3 cases yielded a relative risk of 1.47 for the cluster area. The analysis by zip coded identified the same general area as the town analysis.

At both the town and cluster levels, information is available about the racial/ethnic composition of the communities, the educational attainment levels,

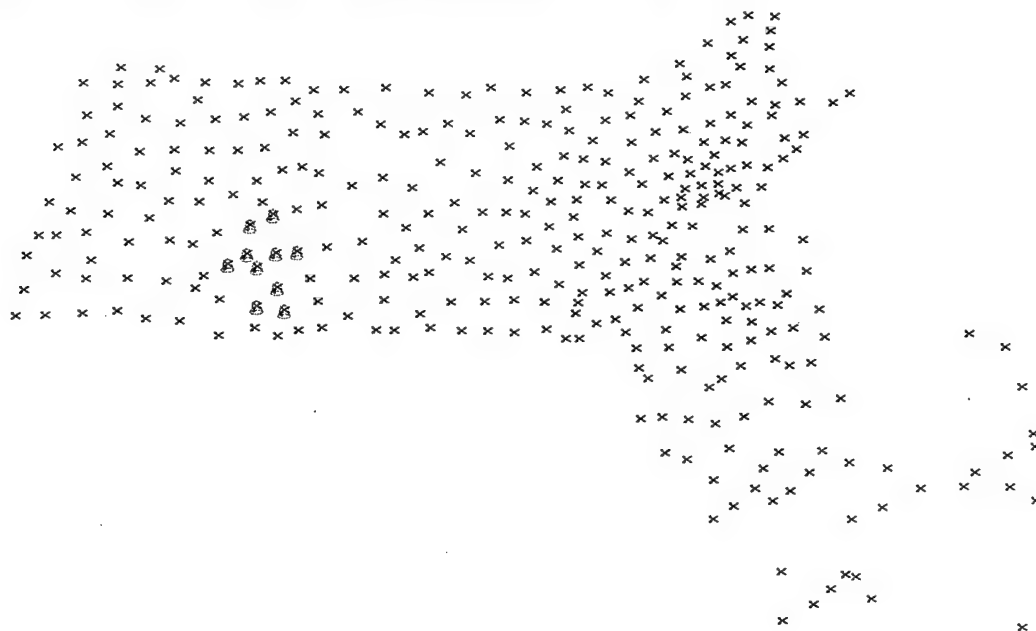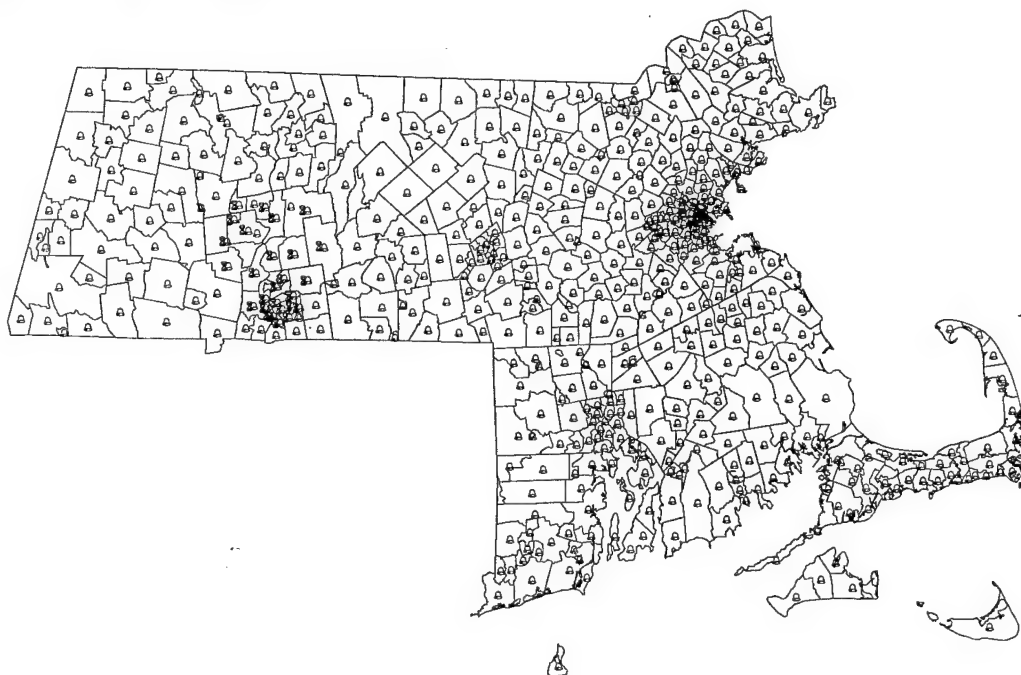Figure 6. Cluster of towns with high proportion of Stage 3 cases.

Figure 7. Cluster of 5 digit zipcode areas with high proportion of Stage 3 cases in dark. Light circles are the zipcode centroids.

unemployment rates, income data, and much more, which may be helpful to policy makers and planners. It is also easy with the GIS system to add other map features, such as road and transportation systems which might facilitate access to screening sites, or rivers and highways, which might impede such access. For example, Figure 8 shows an enlargement of the cluster area and includes the locations of mammography sites. That figure seems to indicate one segment of the cluster area which is not close to a mammography site, an observation which could be validated by persons living or working in that area.

Table 2 contains summary data on the racial/ethnic, educational, and economic factors from the 1990 census limited to Hampshire and Hampden counties, the region in which the cluster towns and zip codes are located. With 39 tracts in the low group and 36 tracts in the high group, all differences are in the expected direction and all differences are statistically significant at $p=.05$, except for the education measures, which have p values of .09 and .18, respectively. It should be pointed out, however, that the observed power is only .40 and .31 for those two comparisons. Furthermore, quite aside from statistical significance, this information should be useful to intervention planners if they are to reach those most at risk. It is also noteworthy that none of these social, educational, or economic measures show significant correlations with the proportion of Stage 1 cases, suggesting perhaps that planners will obtain more information by focusing on the proportion of Stage 3 cases.

Table 2. Low and high Stage 3 tracts within the cluster area (Hampshire and Hampden Counties, Massachusetts) by selected sociodemographic variables from the 1990 Census.						
Proportion Stage 1	< 9 yrs ed	College Grads	Black	Hispanic	Unemployed	Per capita Income
Lowest 42.7%	9.73%	23.86%	4.87%	8.88%	6.86%	$ 14,214
Highest 25%	13.37%	18.01%	12.01%	19.47%	9.93%	$ 11,739

CONCLUSIONS:

This study demonstrates how data from diverse sources can be integrated and analyzed geographically to assess screening efficacy. This system can be used by public health officials to monitor breast cancer screening in particular areas, and should be easily adaptable to monitor other kinds of cancers. In our demonstration we identified a specific geographical area which shows a higher proportion of cases diagnosed at Stage 3 than the rest of the state. Assuming that the same pattern is found with data from 1987 through 1992, those responsible for conducting screening programs within that area might be alerted and provided with further information. This concomitant information about the

Figure 8. Location of mammography units (unfilled symbols) within the cluster of zip codes.
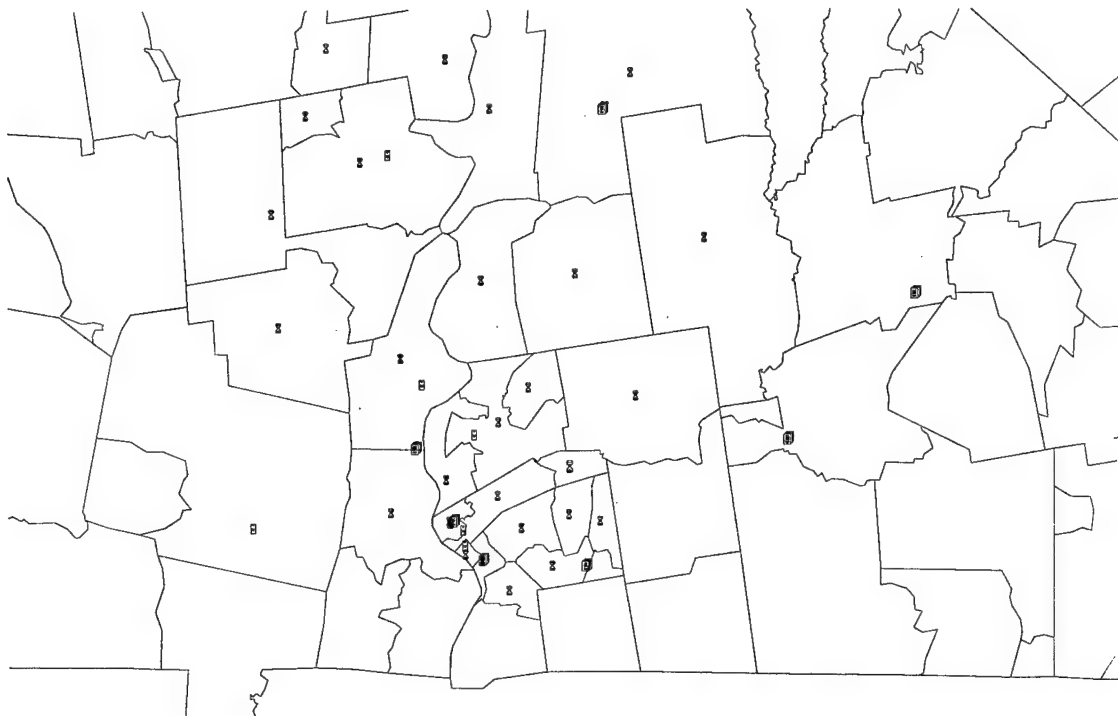
region from the census could be helpful in designing effective interventions. The socioeconomic and racial/ethnic associations with early and late stages of diagnosis are not new (Farley, 1989; Mandelblatt, 1995; Guralnik, 1997); what is new is being able to single out a particular geographical region with statistically significant excesses and immediately access the related socioeconomic and racial/ethnic characteristics of that region and put that information into the hands of intervention planners. It is known that interventions work better if they take target population characteristics into account. A conclusion from the University of Massachusetts Medical Center screening intervention study (Zapka, 1993) recommended "targeted activities aimed at population subgroups" and that the evaluation designs include "broader geographic random samples."

More importantly, while the GIS data system demonstrated here has used breast cancer surveillance as an illustration, it could just as easily have been used to monitor other cancers recorded with the MCR. Furthermore, in this study only proxy measures were available for mammography utilization. Roffers and Austin (1993) suggest that the measure "percent *in situ* of all cases" can reflect frequency of mammography screening and the measure "percent localized of all invasive cases of known stage" may reflect frequency of manual screening. It is entirely possible to incorporate actual mammography utilization data as it becomes available.

Analyses in this study were conducted at the levels of the census tract, the town, and the 5 digit zip code. Because 15.7% of the cases were unusable due to geocoding difficulty at the tract level, that analysis may be the least reliable. Both the analysis at the town level and the zip code level identified the same geographic area as high in the proportion of Stage 3 cases, suggesting that screening programs may need to be improved in that area. It is also possible to conduct evaluations using lower level aggregations, for example, the census block level, as Krieger demonstrated in her San Francisco study (1992). Such finer analyses may be needed in urban areas, while analysis at higher levels of aggregation, such as towns (MCDs), or even CHNAs might be appropriate for certain kinds of studies. The fact that both the analysis by town and zip code revealed the same geographical region should be tempered somewhat by Kulldorff's reminder that "we cannot determine the exact location and shape of any detected cluster, but only the general location."[3]

It should be recognized that cases are assigned to census tracts on the basis of what should be the patient's usual residential address at the time of diagnosis. Problems in the address fields may occur when a business or mailing address is used rather than the residential address, or when a temporary address instead of a usual residence address is used, such as when a patient is staying with a relative or friend during care. All of these examples would introduce errors in assigning census tracts. In addition, the geocoding process itself introduces tracting errors through errors in the GIS data. Examples would include inexact alignment of street-level data overlain on census tract boundaries,

---

[3] M. Kulldorff, personal communication, May 1996.

misnumbered buildings, misnamed streets, inverted block numbering, missing building numbers and street names.

Another caution entails the assignment of tract-level socioeconomic data to individual-level cases within those tracts, as persons diagnosed with breast cancer within a tract may not be typical of other residents within those tracts. Krieger (1992), however, has found that the use of socioeconomic data, at least at the level of the tract and block, is generally not misleading, and is consistent with the findings of others, and probably underestimates the effects that would have been observed were individual-level data available.

Despite these cautions, the overriding importance of screening and the creation of tools to assess its efficacy cannot be emphasized enough. As Marchant's (1994) recent review of contemporary management of breast cancer concluded, "it is undeniable that early detection and treatment of breast cancer reduces morbidity and mortality, and mammography screening is the only method available to detect cancer at the earliest stage when it is most likely to be cured."

REFERENCES

Andrews HF, Kerner JF, Zauber AG, Mandelblatt J, Pittman J and Struening E. Using census and mortality data to target small areas for breast, colorectal, and cervical cancer screening. *American Journal of Public Health.* 1994;84(1):56-61.

Boss LP and Suarez L. Uses of data to plan cancer prevention and control programs. *Public Health Reports.* 1990;105(4):354-360.

Chevarley F and White E. Recent Trends in Breast Cancer Mortality among White and Black US Women. *American Journal of Public Health.* 1997;87(5):775-781.

Chu K, Smart C and Tarone R. Analysis of breast cancer mortality and stage distribution by age for the Health Insurance Plan Clinical Trial. *Journal of the National Cancer Institute.* 1988;80(14):1125-32.

Collette HJ, de Waard F, Rombach JJ, Collette C and Day NE. Further evidence of benefits of a (non-randomized) breast cancer screening programme: the DOM Project. *Journal of Epidemiology & Community Health.* 1992; 46:382-386.

Dangermond J. How to cope with geographical information systems in your organization. In: Scholten HJ and Stilwell JC, editors. Geographic information systems for urban and regional planning. Dordrecht, Kluwer Academic Publication, 1990.

Farley TA and Flannery JT. Late-stage diagnosis of breast cancer in women of lower socioeconomic status: public health implications. *American Journal of Public Health.* 1989;79(11):1508-1512.

Foster RS, Farwell ME and Costanza MC. Breast-conserving surgery for breast cancer: patterns of care in a geographic region and estimation of potential applicability. *Annals of Surgical Oncology.* 1995;2(3):275-280.

Gershman ST, MacDougall LA, Wood MC, Cory J and Palombo R. *Cancer in Massachusetts Women 1982-1994.* Massachusetts Department of Public Health, 1997.

Guralnik JM and Leveille SG. Annotation: Race, ethnicity, and health outcomes -- unraveling the mediating role of socioeconomic status. *American Journal of Public Health.* 1997;87(5):728-730.

Habbema JD, van Oortmarssen GJ, van Putten DJ, Lubbe JT and van der Maas PJ. Age-specific reduction in breast cancer mortality by screening: an analysis of the results of the Health Insurance Plan of Greater New York Study. *Journal of the National Cancer Institute.* 1986;77(2):317-20.

Kerner JF, Struening E, Pittman J, Andrews H, Sampson N and Strickman N. Small area variation in cancer incidence and mortality: a methodology for targeting cancer control programs. In: *Advances in Cancer Control Epidemiology and Research*. New York: Alan R. Liss, 1984:225-234.

Krieger N. Overcoming the absence of socioeconomic data in medical records: validation and application of census-based methodology. *American Journal of Public Health*. 1992;82(5):703-710.

Kulldorff M and Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine*. 1995;14:799-810.

Kulldorff M. Statistical methods for spatial epidemiology: tests for randomness. In: Gatrell and Loytonen (eds). GIS and Health. London: Taylor and Francis, 1997a.

Kulldorff M, Feuer EJ, Miller BA and Freedman LS. Breast cancer clusters in the northeast United States: a geographic analysis. *American Journal of Epidemiology*. 1997b;146(2):161-170.

Lantz P, Bunge M, Cautley E, Phillips JL and Remington PL. Mammography use -- Wisconsin, 1980-1983. *Morbidity and Mortality Weekly Report*. 1995 (October 13):754-756.

Lidbrink E, Frisell J, Brandberg Y, Rosendahl I and Rutqvist L-E. Nonattendance in the Stockholm mammography screening trial: Relative Mortality and reasons for nonattendance. *Breast Cancer Research and Treatment*. 1995;35(3):267-275.

Maguire DH, Goodchild MF and Rhind DW, editors. Geographic information systems, two volumes. Harlow, England: Longman Scientific & Technical, 1991.

Mandelblatt J, Andrews H, Kao R, Wallace R and Kerner J. Impact of Access and Social Context on Breast Cancer Stage at Diagnosis. *Journal of Health Care for the Poor and Underserved*. 1995;6(3):342-351.

Marchant DJ. Contemporary management of breast cancer. *Obstetrics and Gynecology Clinics of North America*. 1994;21(4):555-560.

Roffers SD and Austin DF. Cancer registry data measures for breast and cervical cancer control: definitions, applications, and analyses. *The Abstract*. 1993 (April);13-15.

Shapiro S, Venet W, Strax P, Venet L and Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute*. 1982;69:349-355.

Tabar L, Fagerberg G, Duffy SW, Day NE, Gad A, Grontoft O. Update of the Swedish two-county program of mammographic screening for breast cancer. *Radiologic Clinics of North America*. 1992;30(1):187-210.

Tabar L, Gad A, Holmberg LH, Ljungquist U, Eklund G, Pagerberg CJ, et al. Reduction in mortality from breast cancer after mass screening with mammography. *Lancet*. 1985;1:829-832.

Zapka JG, Costanza ME, Harris DR, Hosmer D, Stoddard A, Barth R et al. Impact of a breast cancer screening community intervention. *Preventive Medicine*. 1993;22(1):34-53.

POPULATION/DEMOGRAPHIC DATA PREPARED
for the
MASSACHUSETTS CANCER REGISTRY: 1980-1995

One of the research objectives of the Massachusetts Cancer Registry (MCR) has been to determine whether or not statistically significant relationships exist between the incidence of breast and/or cervical cancers among Massachusetts women and socio-demographic characteristics of the geographic areas of their residence. This is a description of the population data sets created for use in these analyses.

There are two types of population information: 1. Population counts and estimates, i.e., the numbers of persons by gender, age, and race/ethnicity residing in specified geographic areas. Counts are reported decennially by the Census Bureau for 100 per cent (ideally) of the population; estimates are made for inter- and post-censal years. 2. Demographic information, i.e., social and economic characteristics reflective of life-styles . This information was collected in 1990 from a 16 percent (approximately) sample of the population and then generalized to the whole.

The elemental geographic units are also of two types: 1. Census Tracts (CTs) and 2. Minor Civil Divisions (MCDs). Separate files were created for the CTs and the MCDs, because tracts are not consistently parts of towns. In metropolitan areas, CTs are generally identified with a given MCD, and they can be aggregated to specify neighborhoods or other localities within the MCD. However, in more sparsely settled areas (e.g., western Massachusetts), tracts may be shared by two or more towns.

Census tracts are relatively permanent subdivisions of a county. They are identified by a basic 4-digit number and may have a 2-digit suffix, in which case they are referred to as 6-digit tracts. Tract numbers range from 0001 to 9499.99 and are unique to a county. The numbers 9501 to 9989.99 denote block numbering areas; 9501-9504 are four such areas in Nantucket County. (The suffix .99 denotes a tract populated entirely by persons aboard one or more civilian or military ships.)

Census tracts were delineated by the Census Bureau in the counties of Barnstable, Dukes, and Franklin for the first time in 1990. Some tract numbers already used in Suffolk County (viz., Boston) were

inadvertently assigned to tracts in Barnstable and Franklin counties. In these cases, the county code is a necessary prefix to identify the unique census tract.

Minor civil divisions are the primary political or administrative divisions of a county. In Massachusetts, they are the 351 cities/towns which have locally elected government officials.

Population Counts/Estimates

Population counts for 1980 and 1990 were provided by the U.S. Census Bureau. They were categorized by gender and five-year groups for each Massachusetts city/town. These were the bases for linearly interpolating annual gender/age values per MCD for 1981-1989. Subsequently, estimates of the 1995 populations by gender/age per MCD were provided by the Massachusetts Institute for Social and Economic Research (MISER) at the University of Massachusetts, Amherst. These were used with the 1990 census counts to derive comparable annual interpolations per MCD for 1991-1994. Each record is coded to permit geographic aggregations of data for the Community Health Network Areas (CHNAs), counties, and other collections of MCDs. These 1980-1995 gender/age populations provide the necessary denominators for various age-adjusted rate determinations for cancer incidence/mortality over time among geographic entities.

Census counts of gender/age distributions among Massachusetts census tracts for 1980 and 1990 were also obtained. No inter-censal interpolations were attempted, however; the tract populations are too small to warrant any implied validity or reliability. Furthermore, there are no estimates of 1995 tract populations which are required for any post-censal interpolations.

Cancer incidence data consist of reported diagnoses since 1982. Each confidential report cites the name and address of the patient which identify the city/town and the census tract of residence. Other information includes the date of diagnosis, site and type of cancer and its stage of development.

For the initial tests for statistical association of cancer incidence and demographic factors,

30

age-adjusted rates of breast and cervical cancers were calculated for MCDs. It soon became clear that census tracts could provide a more suitably refined geographic unit of analysis than could whole cities/towns. Socioeconomic and demographic items from the 1990 census were selected on the basis of demonstrated statistical association with different life-styles. These characteristics, as well as selected population data, were obtained for 6-digit CTs. They were then aggregated to 4-digit tracts, inasmuch as the incidence data were coded to no more than four digits. A file of breast cancer incidence data, including stage of development, for 1982-1986 was created. Files of socio-demographic characteristics for (a) MCDs and (b) CTs were each matched with the incidence file to produce two preliminary work files: (1) TWLFSTYL; n = 351 and (2) TRLFSTYL; n = 1,177. The first file contained a record for each MCD, including the few in which no breast cancers had been diagnosed during the 1982-1986 period. The second file consisted of only tracts in which a breast cancer had been reported.

Statistical analyses of the data in the TRLFSTYL file did not reveal any real relationship between the independent socio-demographic variables and the stage of the cancer when it was diagnosed. There were, however, some interesting geographic patterns of tract clusters. The experience in dealing with the 1982-1986 information has been invaluable as preparation for the analyses of the 1987-1994 data. The coding of in situ observations, begun in 1992, is expected to enhance the value of stage classification as a measure of breast cancer detection effectiveness. The quality of geocoding has been markedly improved, also.

It was felt that socio-demographic information could be of particular value in characterizing geographic areas of interest to health planners. Specific target populations could be more readily located geographically; possible problems in communicating effectively might be anticipated and appropriate program modifications considered.

This is the background for the preparation of the population files for future use with the MCR incidence data.

31

The population counts and estimates are in one file, POPS8095.dat; a text description of the file format is in POPS8095.fmt.

Several socio-demographic data thought to be possibly related to communication, accessibility to health facilities, etc., were gleaned from the 1990 census responses. They were downloaded for the MCDs and the 6-digit CTs, respectively. The latter were aggregated to 4-digit tracts for availability when that is the geographic unit of choice. The latitude and longitude of the centroids of the MCDs and the 6-digit CTs, as furnished by MAPINFO, were also added. Land area in square miles was included for MCDs, in the event comparative density measures might be useful.

These files are described in detail in the attached hard copies. They are on disk and in the network's H:\SHARED\POPLNMCR*.* as

A. 1980-1995 populations by gender/age: POPS8095.dat (data): 1,338,624 bytes
                                        POPS8095.fmt (format): 585 bytes

B. 1990 socio-demographic variables
        Towns/cities: MCDVAR97.dat: 184,978 bytes
                      MCDVAR97.fmt: 34,816 bytes

        Tracts: TR6VAR97.dat: 501,221 bytes
                      TR6VAR97.fmt: 28,672 bytes
                      TR4VAR97.dat: 490,946 bytes
                      TR4VAR97.fmt: 47,616 bytes

[N.B. The fmt (format) files are in Microsoft Word, version 6.0]

FILE FORMAT for MCDVAR97.DAT

Initially, selected socio-demographic variables for the 351 minor civil divisions (MCDs) of Massachusetts were taken from the 1990 census conducted by the U.S. Bureau of the Census. They were to be used in a statistical investigation of possible significant relationships with breast cancer incidence data collected by the Massachusetts Cancer Registry (MCR).

Later, several more items per MCD were taken from the census data. This is the format which identifies the variables and their locations in the data set.

V 1. Residence//1-3. This is the 3-digit numeric code for MCDs as used by the Mass. Dept.of Public Health (DPH).

V 2.County//5-6. This is the 2-digit code for the 14 Massachusetts counties as used by the DPH.

V 3.Region//8. This is a code used by the MCR to designate each of the 6 regions of Community Health Network Areas (CHNAs).

V 4. New CHNA//9-10. A new numeric code for the 27 CHNAs.

V 5. Old CHNA//12-13. The 2-digit code for the original 27 CHNAs.

V 6 HSA//15-16. The 2-digit codes for Health Service Areas used by the DPH before CHNAs were adopted.

V 7.LTC//18-20. The 3-digit code for Long Term Care sub-areas of the HSAs.

V 8.FED5//22-26. The 5-digit code devised by the Federal government. The thousands of inhabited places in the U.S. were listed alphabetically and then sequentially numbered. In Massachusetts the MCDs range from 00170 (Abington) to 82525 (Yarmouth).

V 9. FED6//28-33. The 6-digit code was also devised by the federal government. The first 3 digits are odd-numbers assigned to the alphabetically ordered counties; e.g., Barnstable is 001, Middlesex is 017, and Worcester is 027. The next 3 digits have been assigned to the alphabet-ically ordered MCDs within each county in increments of 5,.i.e., 005, 010, 015, etc. Each county has an MCD with the code of 005; but when used with the county code, the resulting 6-digit code uniquely denotes one of the 351 MCDs. (A serial listing of the 351 FED6 codes is the equivalent of an alphabetical listing of MCDs within an alphabetical listing of the counties, an array frequently used by the Bureau of the Census).

V 10. Name//35-50, the name of the MCD of residence.

V 11. County-alpha//52-58, the name of the county of residence.

V 12. Latitude//63-71.The latitude of the centroid of the MCD of residence was supplied by MAPINFO; it is measured in degrees to the sixth decimal place.

V 13. Longitude//73-82. The longitude of the MCD's centroid was also supplied by MAPINFO. (At Boston's latitude, a degree of longitude is approximately 50.8 statute miles; one-millionth of such a degree is about three inches.)

These first thirteen variables are geo-codes. The following are the populations and selected demographic and socio-economic variables taken from the 1990 census.

V 14. Total population (of the MCD of residence)//84-89.

V 15. Non-Hispanic Whites//91-96.

V 16. Non-Hispanic Blacks//98-103.

V 17. Non-Hispanic Other-races//105-110. "Other" races are the aggregation of the other two racial categories, native Americans (Indian, Eskimo, et al.) and Asians.

V 18. Hispanics//112-117. Hispanics are considered an ethnic clasification and can be composed of any of the four racial designations. V 15-V 18 are mutually exclusive and sum to the total population (V14).

V 19. Foreign-born//119-124.
V 20. Persons of age 5 or more years who were living in the same house for 5 years (1985 to 1990)//126-131.
V 21. Persons of age 5 or more years//133-138. These are the ones who were alive in 1985, i.e., the denominator for calculating V20 as a per cent.
V 22. Females separated or divorced//140-145.
V 23. Females of age 15 or more years//147-152. This is the universe of women for whom marital status was determined.

V 24. Persons having 8 or fewer years of formal education//154-159.
V 25. Persons having some high school (grades 9, 10, 11) education//161-166.
V 26. High school graduates and persons who have had some college education//168-173.
V 27. Persons having 4 or more years of college education//175-180.
V 28. Persons of age 25 or more years//182-187. This is the reference population for measuring/comparing levels of educational attainment.

V 29. Unemployed persons//189-194.
V 30. Civilian labor force//196-201. Persons of age 16 or more years, the denominator when calculating unemployment rates.

V 31. Persons whose 1989 income was below the poverty level//203-208. In 1989 the average poverty threshold for a family of four persons was $12,674.
V 32. Persons for whom poverty status was determined//210-215. Some persons are excluded (e.g., those in institutions, in military group quarters, college dormitories, and unrelated individuals under 15 years of age) from the denominator when poverty rates are calculated.

V 33. Persons of age 65 or more years (the "elderly")//217-222.

V 34. Per capita income (received in 1989)//224-229.
V 35. Persons receiving Public Assistance Income//231-236
V 36. Public Assistance denominator//238-243. This is the sum of persons receiving and persons NOT receiving Public Assistance. It should equal the MCD's total population, but these answers were obtained from the sample of census respondents who were sent the 'long' form of the questionnaire. This number is an estimate of the total population based on a sample of approximately 16%.; the total population reported as V 14 (above) is the number in the 100% sample.

V 37. Persons whose 1989 income was less than half the poverty level//245-250; i.e., an income of less than $6,337.
V 38. Persons whose 1989 income was at least twice the poverty level//252-257; i.e., an income of at least $25,348. (The denominator for V 37 and V 38 is V 32, above.)

V 39. Females reporting limitations of mobility and/or self-care//259-264. A series of questions tried to determine the number of persons, by gender and age, who were physically handicapped. The responses were aggregated to a simple summary of YES or NO for each gender. Only females are considered here, because the universe from the Cancer Registry is females for whom breast cancer has been diagnosed.
V 40. The denominator for V 39//266-271. These are the women among only the

non-institutionalized civilian population of age 16 or more years--the universe for the questions about physical limitations.

V 41. Persons in Owner-occupied housing units//273-278. This numerator can be used to determine the proportion of housing units that are Owner-occupied rather than rented; it can be used to calculate the proportion of an MCD's population who own their homes; and it can be used with V 42 to calculate persons-to-unit ratios among Owner-occupied homes.

V 42. Owner-occupied housing units//280-285.

V 43. Persons in Renter-occupied housing units//287-292.

V 44. Renter-occupied housing units//294-299.

V 45. Occupied housing units//301-306. (The sum of V 42 and V 44.)

V 46. Total housing units//308-313. (This minus V 45 is the number of vacant housing units.)

V 47. Water source: Public/Private//315-320.

V 48. Water source: Well//322-327.

V 49. Water source: Other//329-334.

V 50. Sewage disposed: Public sewer//336-341.

V 51. Sewage disposed: Septic/cess pool//343-348.

V 52. Sewage disposed:Other means//350-355.

(The universe for V 47 through V 52 is ALL housing units, V 46.)

V 53. Telephone in Owner/Renter household//357-362. (Universe = Occupied HH, V 45.)

V 54. Zero vehicles per occupied household//364-369.

V 55. Two or more vehicles per occupied household//371-376.

V 56. Kitchen facilities = complete//378-383.

V 57. Plumbing facilities = complete//385-390.

(Universe for V 56 and V 57 = All housing units, V 46.)

V 58. Complete plumbing facilities in WHITE occupied households//392-397.

V 59. WHITE occupied households//399-404.

V 60. Complete plumbing facilities in BLACK occupied households//406-411.

V 61. BLACK occupied households//413-418.

V 62. Complete plumbing facilities in OTHER occupied households//420-425.

V 63. OTHER occupied households//427-432.

V 64. Complete plumbing facilities in HISPANIC occupied households//434-439.

V 65. HISPANIC occupied households//441-446.

Variables 66-73 are concerned with the Census-reported per capita income by race and ethnicity. The race categories are White, Black, and Other (the combined numbers of American Indians and Asians); the ethnic group is the Hispanics. [Note that these Census Bureau's racial groups do not equal the numbers of non-Hispanic Whites, Blacks, and Others, Variables 15-17.]

V 66. Whites//448-453.

V 67. WHITE per capita income//455-460.

V 68. Blacks//462-467.

V 69. BLACK per capita income//469-474.

V 70. Others//476-481.

V 71. OTHER per capita income//483-488.

V 72. Hispanics//490-495. (Same as V 18, above.)

V 73. HISPANIC per capita income//497-502.

V 74. Land area//504-509. This variable is expressed in square miles to two decimal places. It is the denominator for measures of density (e.g., population, breast cancer cases) for a

given MCD. It is additive, i.e., the land area of a CHNA can be determined by summing the land areas of the constituent MCDs.

V 75. Water area//511-516. This complements land area to give the TOTAL area of an MCD in square miles.

V 76. Population density//518-525. This is the ratio of the total population (V 14) to the land area expressed in square miles (V 74).

FILE FORMAT for TR6VAR97.DAT

Initially, selected socio-demographic variables for the 1,331 census tracts (CTs) of Massachusetts were taken from the 1990 census conducted by the U.S. Bureau of the Census. They were to be used in a statistical investigation of possible significant relationships with breast cancer incidence data collected by the Massachusetts Cancer Registry (MCR).

Later, several more items per CT were taken from the census data. This is the format which identifies the variables and their locations in the data set. Twelve census tracts had zero population and were not included in this data set of 1,319 records.

V 1. County//1-3. This is the 3-digit numeric code for counties as used by the Bureau of the Census. They are the odd-numbers assigned to an alphabetical listing of the fourteen Massachusetts counties: 001 = Barnstable,...,027 = Worcester.

V 2.Tract//4-9. This is the 6-digit code for the 1,331 Massachusetts tracts.

There are eight tracts in both Barnstable and Suffolk counties which have the same codes; similarly, there are eight tracts in both Franklin and Suffolk counties which have the same codes. Since the county is necessary to identify these 16 tracts, the usual tract identifier is the 9-digit field 1-9. (This field can be used also as an edit-check, because among the other 1,303 tracts each county has a unique range of tract codes; any tract code not within the county's specific range is incorrect.)

V 3. County-alpha//11-20. This is a the name of the Massachusetts county.

V 4. Region//22. This is the code used by the MCR to designate each of the six regions of the Community Health Network Areas (CHNAs).

V 5. New CHNA//23-24. A new numeric code for the 27 CHNAs.

V 6. Old CHNA//26-27. The 2-digit code for the original 27 CHNAs.

V 7. Latitude//29-37.The latitude of the centroid of the tract was supplied by MAPINFO; it is measured in degrees to the sixth decimal place.

V 8. Longitude//39-48. The longitude of the tract's centroid was also supplied by MAPINFO.

The above variables are geo-codes. The following are the populations and selected demographic and socio-economic variables taken from the 1990 census.

V 9. Total population (of the CT of residence)//50-54

V 10. Non-Hispanic Whites//56-60.

V 11. Non-Hispanic Blacks//62-66.

V 12. Non-Hispanic Other-races//68-72. "Other" races are the aggregation of the other two racial categories, native Americans (Indian, Eskimo, et al.) and Asians.

V 13. Hispanics//74-78. Hispanics are considered an ethnic clasification and can be composed of any of the four racial designations. V 10-V 13 are mutually exclusive and sum to the total population (V 9).

V 14. Foreign-born//80-84.

V 15. Persons of age 5 or more years who were living in the same house for 5 years (1985 to 1990)//86-90.

V 16. Persons of age 5 or more years//92-96. These are the ones who were alive in 1985, i.e., the denominator for calculating V 15 as a per cent.

V 17. Females separated or divorced//98-102.

V 18. Females of age 15 or more years//104-108. This is the universe of women for whom marital status was determined.

V 19. Persons having 8 or fewer years of formal education//110-114.

V 20. Persons having some high school (grades 9, 10, 11) education//116-120.

V 21. High school graduates and persons who have had some college education//122-126.

V 22. Persons having 4 or more years of college education//128-132.

V 23. Persons of age 25 or more years//134-138. This is the reference population for measuring/comparing levels of educational attainment.

V 24. Unemployed persons//140-144.

V 25. Civilian labor force//146-150. Persons of age 16 or more years, the denominator when calculating unemployment rates.

V 26. Persons whose 1989 income was below the poverty level//152-156. In 1989 the average poverty threshold for a family of four persons was $12,674.

V 27. Persons for whom poverty status was determined//158-162. Some persons are excluded (e.g., those in institutions, in military group quarters, college dormitories, and unrelated individuals under 15 years of age) from the denominator when poverty rates are calculated.

V 28. Persons of age 65 or more years (the "elderly")//164-168.

V 29. Per capita income (received in 1989)//170-174.
(Please note that the locations of the next three income variables are out of sequence.)

V 30. Per capita income of Whites//188-192.

V 31. Per capita income of Blacks//194-198.

V 32. Per capita income of Hispanics//200-204.

V 33. Persons receiving Public Assistance Income//176-180

V 34. Public Assistance denominator//182-186. This is the sum of persons receiving and persons NOT receiving Public Assistance. It should equal the CT's total population, but these answers were obtained from the sample of census respondents who were sent the 'long' form of the questionnaire. This number is an estimate of the total population based on a sample of approximately 16%.; the total population reported as Variable 9 (above) is the number in the 100% sample.

V 35. Persons whose 1989 income was less than half the poverty level//206-210; i.e., an income of less than $6,337.

V 36. Persons whose 1989 income was at least twice the poverty level//212-216; i.e., an income of at least $25,348. (The denominator for V 35 and V 36 is V 27, above.)

V 37. Females reporting limitations of mobility and/or self-care//218-222. A series of questions tried to determine the number of persons, by gender and age, who were physically handicapped. The responses were aggregated to a simple summary of YES or NO for each gender. Only females are considered here, because the universe from the Cancer Registry is females for whom breast cancer has been diagnosed.

V 38. The denominator for V 37//224-228. These are the women among only the non-institutionalized civilian population of age 16 or more years--the universe for the questions about physical limitations.

V 39. Persons in Owner-occupied housing units//230-234. This numerator can be used to determine the proportion of housing units that are Owner-occupied rather than rented; it can be used to calculate the proportion of a CT's population who own their homes; and it can be used with V 40 to calculate persons-to-unit ratios among Owner-occupied homes.

V 40. Owner-occupied housing units//236-240.

V 41. Persons in Renter-occupied housing units//242-246.

V 42. Renter-occupied housing units//248-252.
V 43. Occupied housing units//254-258. (The sum of V 40 and V 42.)
V 44. Total housing units//260-264. (This minus V 43 is the number of vacant housing
units.)

V 45. Water source: Public/Private//266-270.
V 46. Water source: Well//272-276.
V 47. Water source: Other//278-282.
V 48. Sewage disposal: Public sewer//284-288.
V 49. Sewage disposal: Septic/cess pool//290-294.
V 50. Sewage disposal:Other means//296-300.
      (The universe for V 45 through V 50 is ALL housing units, V 44.)

V 51. Telephone in Owner/Renter household//302-306. (Universe = Occupied HH, V 43.)
V 52. Zero vehicles per occupied household//308-312.
V 53. Two or more vehicles per occupied household//314-318.

V 54. Kitchen facilities = complete//320-324.
V 55. Plumbing facilities = complete//326-330.
      (Universe for V 54 and V 55 = ALL housing units, V 44.)

V 56. Complete plumbing facilities in WHITE occupied households//332-336.
V 57. WHITE occupied households//338-342.
V 58. Complete plumbing facilities in BLACK occupied households//344-348.
V 59. BLACK occupied households//350-354.
V 60. Complete plumbing facilities in OTHER occupied households//356-360.
V 61. OTHER occupied households//362-366.
V 62. Complete plumbing facilities in HISPANIC occupied households//368-372.
V 63. HISPANIC occupied households//374-378.

FILE FORMAT for TR4VAR97.DAT

Initially, selected socio-demographic variables for the 1,331 6-digit census tracts (CTs) of Massachusetts were taken from the 1990 census conducted by the U.S. Bureau of the Census. They were to be used in a statistical investigation of possible significant relationships with breast cancer incidence data collected by the Massachusetts Cancer Registry (MCR). The smallest geographic entity used for the incidence data was the 4-digit census tract. Consequently, the census information for the 6-digit tracts was aggregated to 4-digit tracts. Eight of these were excluded because they were populated only by persons aboard civilian or military ships. The remaining 1,183 constitute the file of 4-digit census tracts.

V 1. County//1-3. This is the 3-digit numeric code for counties as used by the Bureau of the Census. They are the odd-numbers assigned to an alphabetical listing of the fourteen Massachusetts counties: 001 = Barnstable,...,027 = Worcester.

V 2. Tract//4-7. This is the 4-digit code for the 1,183 Massachusetts tracts in the data set.

There are eight tracts in both Barnstable and Suffolk counties which have the same codes; similarly, there are eight tracts in both Franklin and Suffolk counties which have the same codes. Since the county is necessary to identify these 16 tracts, the usual tract identifier is the 7-digit field 1-7.

V 3. County-alpha//9-18. This is a the name of the Massachusetts county.

V 4. Region//20. This is the code used by the MCR to designate each of the six regions of the Community Health Network Areas (CHNAs).

V 5. New CHNA//21-22. A new numeric code for the 27 CHNAs.

V 6. Old CHNA//24-25. The 2-digit code for the original 27 CHNAs.

V 7. Latitude//27-35. The latitude of the centroid of the tract was supplied by MAPINFO; it is measured in degrees to the sixth decimal place.

V 8. Longitude//37-46. The longitude of the tract's centroid was also supplied by MAPINFO. (At Boston's latitude, a degree of longitude is approximately 50.8 statute miles; one-millionth of such a degree is about three inches.)

The above variables are geo-codes. The following are the populations and selected demographic and socio-economic variables taken from the 1990 census.

V 9. Total population (of the CT of residence)//49-53

V 10. Non-Hispanic Whites//55-59

V 11. Non-Hispanic Blacks//61-65.

V 12. Non-Hispanic Other-races//67-71. "Other" races are the aggregation of the other two racial categories, native Americans (Indian, Eskimo, et al.) and Asians.

V 13. Hispanics//73-77. Hispanics are considered an ethnic clasification and can include persons of any of the four racial designations. V 10-V 13 are mutually exclusive and sum to the total population (V 9).

V 14. Foreign-born//79-83.

V 15. Persons of age 5 or more years who were living in the same house for 5 years (1985 to 1990)//85-89.

V 16. Persons of age 5 or more years//91-95. These are the ones who were alive in 1985, i.e., the denominator for calculating V 15 as a per cent.

V 17. Females separated or divorced//97-101.

V 18. Females of age 15 or more years//103-107. This is the universe of women for whom marital status was determined.

V 19. Persons having 8 or fewer years of formal education//109-113.

V 20. Persons having some high school (grades 9, 10, 11) education//115-119.

V 21. High school graduates and persons who have had some college education//121-125.

V 22. Persons having 4 or more years of college education//127-131.

V 23. Persons of age 25 or more years//133-137. This is the reference population for measuring/comparing levels of educational attainment.

V 24. Unemployed persons//139-143.

V 25. Civilian labor force//145-149. Persons of age 16 or more years, the denominator when calculating unemployment rates.

V 26. Persons whose 1989 income was below the poverty level//151-155. In 1989 the average poverty threshold for a family of four persons was $12,674.

V 27. Persons for whom poverty status was determined//157-161. Some persons are excluded (e.g., those in institutions, in military group quarters, college dormitories, and unrelated individuals under 15 years of age) from the denominator when poverty rates are calculated.

V 28. Persons of age 65 or more years (the "elderly")//163-167.

V 29. Persons receiving Public Assistance Income//169-173

V 30. Public Assistance denominator//175-179. This is the sum of persons receiving and persons NOT receiving Public Assistance. It should equal the CT's total population, but these answers were obtained from the sample of census respondents who were sent the 'long' form of the questionnaire. This number is an estimate of the total population based on a sample of approximately 16%.; the total population reported as Variable 9 (above) is the number in the 100% sample.

V 31. Persons whose 1989 income was less than half the poverty level//181-185; i.e., an income of less than $6,337.

V 32. Persons whose 1989 income was at least twice the poverty level//187-191; i.e., an income of at least $25,348. (The denominator for V 31 and V 32 is V 27, above.)

V 33. Females reporting limitations of mobility and/or self-care//193-197. A series of questions tried to determine the number of persons, by gender and age, who were physically handicapped. The responses were aggregated to a simple summary of YES or NO for each gender. Only females are considered here, because the universe from the Cancer Registry is females for whom breast cancer has been diagnosed.

V 34. The denominator for V 33//199-203. These are the women among only the non-institutionalized civilian population of age 16 or more years--the universe for the questions about physical limitations.

V 35. Persons in Owner-occupied housing units//205-209. This numerator can be used to determine the proportion of housing units that are Owner-occupied rather than rented; it can be used to calculate the proportion of a CT's population who own their homes; and it can be used with V 36 to calculate persons-to-unit ratios among Owner-occupied homes.

V 36. Owner-occupied housing units//211-215.

V 37. Persons in Renter-occupied housing units//217-221.

V 38. Renter-occupied housing units//223-227.

V 39. Occupied housing units//229-233. (The sum of V 36 and V 38.)

V 40. Total housing units//235-239. (This minus V 39 is the number of vacant housing units.)

V 41. Water source: Public/Private//241-245.
V 42. Water source: Well//247-251.
V 43. Water source: Other//253-257.
V 44. Sewage disposal: Public sewer//259-263.
V 45. Sewage disposal: Septic/cess pool//265-269.
V 46. Sewage disposal:Other means//271-275.
      (The universe for V 41 through V 46 is ALL housing units, V 40.)

V 47. Telephone in Owner/Renter household//277-281. (Universe = Occupied HH; V 39.)
V 48. Zero vehicles per occupied household//283-287.
V 49. Two or more vehicles per occupied household//289-293.

V 50. Kitchen facilities = complete//295-299.
V 51. Plumbing facilities = complete//301-305.
      (Universe for V 50 and V 51 = ALL housing units, V 40.)

V 52. Complete plumbing facilities in WHITE occupied households//307-311.
V 53. WHITE occupied households//313-317.
V 54. Complete plumbing facilities in BLACK occupied households//319-323.
V 55. BLACK occupied households//325-329.
V 56. Complete plumbing facilities in OTHER o`ccupied households//331-335.
V 57. OTHER occupied households//337-341.
V 58. Complete plumbing facilities in HISPANIC occupied households//343-347.
V 59. HISPANIC occupied households//349-353.

V 60. Total population//355-359. (Same as V 9: 49-53.)
V 61. Per capita income of TOTAL population//361-365.

The Bureau of the Census reported per capita income for the four major racial groups, each of which may include persons of Hispanic ethnicity. Thus, the numbers of persons in the racial groups of variables 62, 64, and 66 are different from those in variables 10, 11, and 12.

V 62. WHITE population//367-371.
V 63. WHITE per capita income//373-377.
V 64. BLACK population//379-383.
V 65. BLACK per capita income//385-389.
V 66. OTHER population// 391-395.
V 67. OTHER per capita income//397-401.
V 68. HISPANIC population//403-407.
V 69. HISPANIC per capita income//409-413.

Adhoc Windows Application - Adhoc
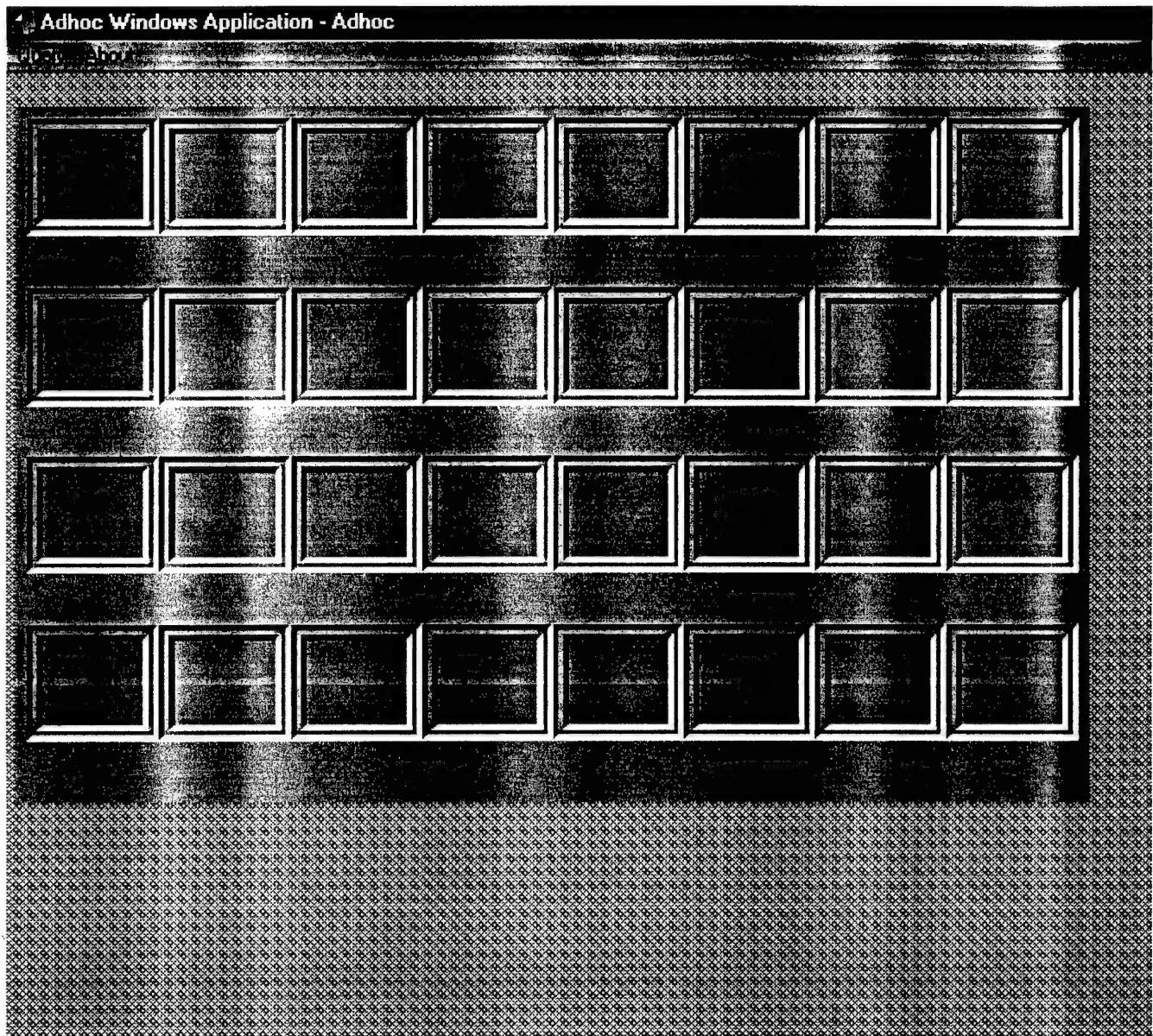
Query   About

Construct Adhoc Query   ☒

Geograhic Unit	Hospital	Diagnosis Date	Gender
Race	Smoking	Age	Cancer
Industrial Code	Occupation Code		

NEW REGULAR QUERY

QUERY NAME : Untitle

Define Report   Save   Save As   Execute   Clear   Delete   Close

Geographic Unit Selection

Geographic Units

Non Group Items

- ○ State
- ○ City / Town
- ○ County
- ○ Census Tract
- ○ CHNA
- ○ ATSDR Sites
- ○ Zip Code

GROUP SELECTION

Group List

Group Name

Clear Group

Group Items

Add Group

Delete Group

☐ Group All Other Items

Delete All   OK   Cancel

Define Report   Save   Save As   Execute   Clear   Delete   Close

Query About

Construct Adhoc Query

Hopsital Selection

Addison Gilbert
Amesbury Hospital
Anna Jacques
Athol Hospital
AtlantiCare Medical Center
Bay State Medical

Non Group Items

QU

GROUP SELECTION

Group List

Group Name

Clear Group

Group Items

Add Group

Delete Group

☐ Group All Other Items

Delete All

Cancel

| Define Report | Save | Save As | Execute | Clear | Delete | Close |

Query About

Construct Adhoc Query

Geog **Date of Diagnosis**

Select by Dates

From [ ] [15] To [ ] [15] **Add**

Indus

Select by Years | Selected Date Ranges

QUERY NAME

1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993

Select | **Delete**

Delete All | **OK** | **Cancel**

Define Report | **Save** | **Save As** | **Execute** | **Clear** | **Delete** | **Close**

Query About

Construct Adhoc Query ☒

Race Selection ☒

American Indian, Aleutian or Esk
Asian Indian, Pakistani
Black
Chamorran
Chinese
Fiji Islander

Non Group Items

QU

GROUP SELECTION

Group List

Group Name

Clear Group

Group Items

Add Group

Delete Group

☐ Group All Other Items

Delete All

| Define Report | Save | Save As | Execute | Clear | Delete | Close |

48

Adhoc Windows Application - Adhoc

Query   About

Construct Adhoc Query

G Age Selection

In

QUERY N

16 Commonly used Age Groups

☐	0 - 4	☒	45 - 49
☐	5 - 9	☐	50 - 54
☐	10 - 14	☐	55 - 59
☐	15 - 19	☐	60 - 64
☐	20 - 24	☐	65 - 69
☐	25 - 29	☐	70 - 74
☐	30 - 34	☐	75 - 79
☐	35_39	☐	80 - 84
☒	40 - 44	☐	85+

Select All   Unselect All

Define Custom Age Groups

From [   ]   To [   ]

Add

Selected Age Groups

40 - 44
45 - 49

Delete

Delete All   OK   Cancel

Define Report   Save   Save As   Execute   Clear   Delete   Close

Adhoc Windows Application - Adhoc

Query About

Construct Adhoc Query

Cancer Selection

Histology Selection

Code

- 800-899
- 900-975
- 980-994
- 995-997
- 999

Associated Term

[Include] [Exclude]

Behavior Code

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4
☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9

Primary Selection

Code

- C44
- C47
- C48
- C49
- **C50**
- C51-C58
- C60-C63

Associated Term

Breast (excludes skin of breast C44.5)

[Include] [Exclude]

Define Custom Range

From [____] To [____]

[Include]
[Exclude]

☐ Group All Others

Cancer List

[_____] ▼

[Clear Cancer] [Add Cancer] [Delete Cancer]

Cancer Name

[_____]

Inclusion List

[Delete]

Exclusion List

[Delete]

[Delete All]   [OK]   [Cancel]

50

A User's Guide

for the Adhoc Application
of the MCR - CIMS and MCR / DoD BCCES

(Massachusetts Cancer Registry - Cancer Information Management System
and Massachusetts Cancer Registry / Department of Defense
Breast Cancer Control Evaluation System)

Major Contents

INTRODUCTION

For simplicity I refer throughout the Guide to the MCR - CIMS, but the instructions are also applicable to the MCR / DoD BCCES.

Most of the windows, dialog boxes and other features in CIMS behave like the corresponding features in Microsoft Office applications. I have tried to identify places in CIMS where features behave in more unexpected ways. You can switch between CIMS and your other Windows applications as usual, but you may find that your PC's free memory limits what you can do while CIMS is running.

When there are several ways to perform a procedure, I usually only provide instructions on one or two alternatives. (I sometimes include one or two alternative keystrokes in square brackets, but many things in CIMS can only be "moused".)

As you use the Adhoc Application, many dialog boxes of different sizes may tend to pile up onscreen. Since you can seldom move/resize these dialogs, you must interact with the dialog that is topmost at any moment. Even if you can see part of an underlying dialog, you can't make it the active window by clicking somewhere in it. To get down to an underlying dialog box, you must first close each box on top of it.

GETTING STARTED

Run the Adhoc Application from your Windows Program Manager. (The procedure for doing this may change. Right now, you must start the database server and tables server before you run the Adhoc. In the future, the necessary servers may start automatically when you choose to run the Adhoc.)

The "Window Full of Frames"

The first thing you should see is a window entitled *Adhoc Windows Application - Adhoc* with 32 empty "frames" (for holding information on the status of data queries). The menu bar contains menus **Query** and **About**. For my own convenience, I refer to this screen throughout the Guide as the "window full of frames".
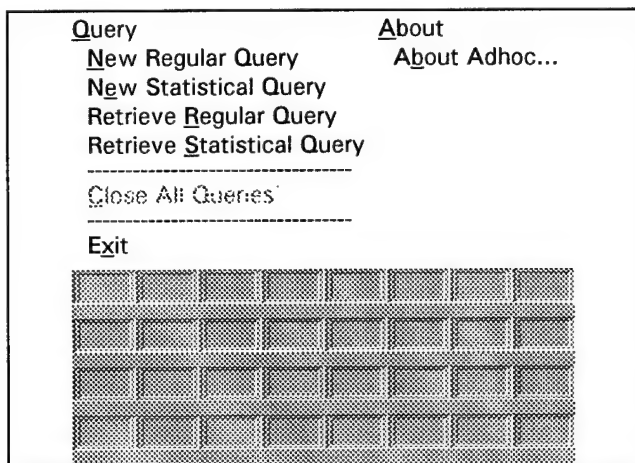
menus and commands in the "window full of frames"

The Query Menu [Alt + Q]

This menu contains 6 commands:

New Regular Query Choose this [N] to build a new data query involving observed case counts only.

New Statistical Query Click here [E] to build a new statistical query involving observed counts, expected counts, confidence intervals, SIRs, age-adjusted incidence rates and/or age-specific incidence rates.

Retrieve Regular Query Click here [R] to choose an observed count query definition you've saved.

Retrieve Statistical Query Click here [S] to choose a statistical query definition you've saved.

Close All Queries This command [C] is available whenever you have a query open.

Exit (to leave the Adhoc Application) [X]

The About Menu

This menu contains a single command which can be used to check the version of the Adhoc you're using. If you should want to do this, choose **About -- About Adhoc** [Alt + A, Enter]. You'll see an information dialog box containing the application's name and version number. To close this dialog, use its control-menu box or click the **OK** button [Esc or Enter]. You will return to the window full of frames.

REGULAR QUERIES

"Regular" data queries search the records in the CIMS database for certain values appearing in certain data fields. You specify the values of the data fields you're interested in as you define the query. When you run a regular query, CIMS produces observed case counts from the records which it found meeting the criteria you specified. For example, you may define a regular query to answer the question, "How many 1988 breast cancer cases do we have for Hispanic Bristol County females over the age of 85?" The 8 data fields we can now "query on" for regular queries are: place of residence, reporting facility, date of diagnosis, sex, race, smoking status, age at diagnosis, and type of cancer. In the future, we should also be able to query on the occupation and industry fields.

Defining a New Regular Query

To define a new data query, from the window full of frames, choose **Query -- New Regular Query** [Alt + Q, N]. You will see a dialog box entitled *Construct Adhoc Query* with a group of command buttons near its top representing the data fields whose values you may specify.

Below these buttons is a text box entitled NEW REGULAR QUERY with scroll bars in which the text summarizing your query definition will appear as you work.

At the bottom of the dialog is a line of command buttons used for query definitions you've started or finished.

Specifying Data Field Values

In order to define a regular query, you must choose and specify values for <u>at least 2</u> of the available data fields. The following instructions describe how to specify values for each field.

Geographic Unit

When you click the **Geographic Unit** button in the *Construct Adhoc Query* dialog box, a dialog entitled *Geographic Unit Selection* appears.

<u>Choose one and only one type of geographic unit</u>[1] from the vertical list under the huge arrowhead by clicking on it [Tab until the **State** option button is blackened, then press ↓ until the option button you want is blackened].

To illustrate how this dialog works, I will explain how to choose towns. The same procedures can be used for any of the other types of (non-statewide) geographic units.

City/Town Example:

If you choose **City/Town** for your type of geographic unit, an alphabetic list of our 351 minor civil divisions appears. Once you've selected any town name in this list, you can move through the list in the usual ways (scroll bars, ↓, ↑, PageUp, PageDown, Home, End). To jump to the first town name beginning with a particular letter, type that letter. (If a town name beginning with that letter is already highlighted, you will move to the next town down).

Selecting town names can be done in the usual "Windows ways". To select consecutive town names, select one of the town names and then Shift + ↑ (to select the next name above), or Shift + ↓, or Shift + PageUp, or Shift + PageDown, or Shift + Home (to select from Abington through the town you initially had selected), or Shift + End (to select Yarmouth through the town you had selected). Another way to select a range of consecutive town names is to select the town name at one end of the range and Shift + click the town name at the other end. To select more than one non-consecutive town name, select one town and then Ctrl + click the other(s).

When you have at least one town name selected, <u>drag</u> (the cursor will become shaped like a little hand) your selection into *either* the **Non-Group Items** box (if you want data for each town separately) *or* into the **Group Items** box (for aggregate data). (A town name can appear in only

---

[1] A query can have *only one type* of geographic unit specified. That is, you cannot have Boston observed counts and statewide observed counts appear in the same query data table. You would have to define 2 different queries for this.

55

one of these 2 boxes at a time.[2] Also, a town name cannot appear in two different groups within one query.[3] You may place a single town name in the **Group Items** box and treat that town as a "group" if you wish.) Any town name(s) you drag away will disappear from the "original list".

You can select town names within the **Non-Group Items** and **Group Items** boxes in the same ways as in the "original list". You can drag selected town names between the **Non-Group Items** and **Group Items** boxes, or back into the "original list".

In the **GROUP SELECTION** area, when you have gotten towns that you want to aggregate into the **Group Items** box, you <u>must</u> give this group some name (up to any 40 characters) by typing it into the **Group Name** box. Then click the **Add Group** button [Tab until it's outlined, then Enter] and the group name you typed will move into the **Group List** drop-down list box. (Drop this list down to see any group names you've created.) You have now finished grouping those towns that were in the **Group Items** box. Repeat these steps until all the town groups you want have been created.

To modify the towns within a group you've created, choose that group's name from the **Group List** box and drag town names into or out of the **Group Items** box as you wish; then click the **Add Group** button again; you'll be asked "Override Existing Group?", so click either **OK** [Enter] to redefine the towns making up that group or **Cancel** [Esc] if you've changed your mind.

To eliminate a group you've created, choose its name from the **Group List** box and click the **Delete Group** button[4]; (you'll be asked a verifying question, so click **OK** [Enter] or **Cancel** [Esc]). If you answer the question with **OK**, look in the **Group List** box and you'll find that that group no longer exists; the town names that were within that group have returned to the "original list".

To change the name you've given to a group, choose its name from the **Group List** box, edit its name in the **Group Name** box (or type a completely new name) and click the **Add Group** button again; say **OK** or **Cancel** when the verifying question appears. You have replaced the old group *name* with a new one, but the towns making up that group have *not* changed.

Clicking the **Clear Group** button (you get a verifying question) just erases the boxes in the **GROUP SELECTION** area so that you can start defining a new group on a "clean slate". The **Clear Group** button does *not* delete or change the makeup of any group you've named and created.

You can click the **Group All Other Items** check box whenever you have at least one town name in the **Non-Group Items** box (regardless of whether or not you have created any groups). If you use this feature, you do not give this group a name, the towns making up the group do not appear in the **Group Items** box, and the group does not appear in the **Group List** box. In essence, this feature aggregates all the town names remaining in the "original list". (Actually, CIMS just calculates the difference between the total observed case count in the database and the total case counts for the towns you've specified.)

When you've finished specifying the geographic unit(s) you want in this query, click the **OK** button at the bottom of the dialog box and you will return to the *Construct Adhoc Query* dialog. Any geographic unit(s) you chose appear in the NEW REGULAR QUERY text box *in alphabetical or numerical order* (regardless of the order in which you chose them). To modify these choices, click the **Geographic Unit** button again and make the necessary changes in the *Geographic Unit Selection* dialog box.

To eliminate everything you've specified in the *Geographic Unit Selection* dialog box, click the **Delete All** button at the bottom of the dialog box. (You will have a chance to change your mind.)

---

[2] That is, you cannot get counts for Fall River individually and aggregate counts for the group {Fall River + Freetown + Somerset} from the same query.

[3] That is, you cannot create groups {Boston + Cambridge} and {Boston + Brookline} in the same query.

[4] <u>NOT</u> the **Delete All** button at the bottom of the dialog!

The **Cancel** button [Esc] (or closing the *Geographic Unit Selection* dialog box using its control-menu box) [Alt+F4] will return you to the *Construct Adhoc Query* dialog box as it was at the time you last clicked the **Geographic Unit** button.

Non-City/Town Geographic Units

If you choose a different type of geographic unit (county, census tract[5], CHNA or ZIP Code), the selection dialog box works in the same way as for towns, except that for numerical units, typing a number will move you to that part of the "original list". The choice of ATSDR sites is not yet available.

If you specify *no* geographic unit when defining a query, this is tantamount to a statewide query. There is no need to specify the option "State" in a query definition, unless you want the word "State" to appear in the query text (as a reminder that the query is for statewide data).

Hospital

To specify a choice of reporting hospital, click the **Hospital** button in the *Construct Adhoc Query* dialog box, and the *Hospital Selection* dialog appears. This dialog works just like the *Geographic Unit Selection* dialog, except that no "type of hospital" needs to be specified.

Diagnosis Date

To specify dates of diagnosis for your query, click the **Diagnosis Date** button in the *Construct Adhoc Query* dialog box, and the *Date of Diagnosis* dialog appears.

```
┌─────────────────────────────────────────────────────────────────────────┐
│ ┌───┐                    Date of Diagnosis                              │
│ │ ─ │                                                                    │
│ └───┘                                                                    │
│   ┌──────────────────────────────────────────────────────────────┐      │
│   │                     Select by Dates                            │      │
│   │  From ┌──────────┐ 15   To ┌──────────┐ 15    ┌─────────┐     │      │
│   │       └──────────┘         └──────────┘       │   Add   │     │      │
│   │                                               └─────────┘     │      │
│   └──────────────────────────────────────────────────────────────┘      │
│   ┌──────────────┐    ┌────────────────────────────────────────┐        │
│   │Select by Years│    │          Selected Date Ranges          │        │
│   │  ┌─────────┐ │    │                                         │        │
│   │  │1982   ↑ │ │    │                                         │        │
│   │  │1983     │ │    │                                         │        │
│   │  │1984     │ │    │                                         │        │
│   │  │1985     │ │    │                                         │        │
│   │  │1986     │ │    │                                         │        │
│   │  │etc.     │ │    │                                         │        │
│   │  │       ↓ │ │    │                                         │        │
│   │  └─────────┘ │    └────────────────────────────────────────┘        │
│   │ ┌──────────┐ │              ┌──────────┐                             │
│   │ │  Select  │ │              │  Delete  │                             │
│   │ └──────────┘ │              └──────────┘                             │
│   └──────────────┘                                                       │
│ the "original list"   ┌───────────┐ ┌──────────┐ ┌──────────┐           │
│                       │ Delete All│ │    OK    │ │  Cancel  │           │
│                       └───────────┘ └──────────┘ └──────────┘           │
└─────────────────────────────────────────────────────────────────────────┘
```

To choose an individual entire year(s) of data, use the **Select by Years** area. Select the year(s) you want and click the **Select** button. The years selected will appear in numerical order[6] in the **Selected Date Ranges** box and will disappear from the original list. You can select several years at a time by

---

[5] These are the 1990 6-digit tracts in numeric order <u>without</u> their county code numbers! Remember that tract numbers are unique within counties, but not within the state. For tract numbers which are duplicates (between Suffolk County and Franklin/Barnstable Counties), I do not know how we'll be able to specify which tract we want.

[6] The order in which you choose dates may be important to you. The order in which you choose dates will be reflected in the query text and in the resulting data tables when the query is run. That is, if you want 1992 counts to appear first in a data table, choose that year first in the *Date of Diagnosis* dialog box.

using Shift+click, Ctrl+click, Shift+↑, Shift+↓, Shift+Home, Shift+End, Shift+PageUp or Shift+PageDown.

To remove a year(s) from the **Selected Date Ranges** box, select it(them) and click the **Delete** button beneath the box[7]; the year(s) you deleted will reappear in the original list.

To aggregate *entire* years of data, use the **Select by Dates** area. (See instructions immediately below.)

If you do not want to choose *entire* years of data (January 1 - December 31), use the **Select by Dates** area. You may type into the **From** and **To** boxes the starting and ending dates of the range you want (in MM/DD/YYYY format), or by clicking on the little calendar (15) in either box and clicking on a date from the perpetual calendar that appears. (These calendars always start by displaying the current date (1996), so use the arrowhead buttons at the top of the calendar screen to change the year shown, or type the year you want where the "1996" appears.) Click on the date in the year you want to start or end with, and click the **OK** button on the right. Click the **Cancel** button to return to the *Date of Diagnosis* dialog box as you left it.) Once you've entered a date into the **Select by Dates** area (but not yet "added" it), it's difficult to "erase" it by backspacing or using the Delete key on the keyboard; so click that 15 and click the **Blank Date** button, then click **OK**. When you have gotten the correct dates in the **From** and **To** boxes, click the **Add** button to their right; and the date range you specified will appear in the **Selected Date Ranges** box below.

When you have all the dates you want in the **Selected Date Ranges** box, click the **OK** button at the bottom of the dialog box. You will return to the *Construct Adhoc Query* dialog box.

To clear all the boxes in the *Date of Diagnosis* dialog box, click the **Delete All** button at the bottom of the dialog (you can change your mind).

To return to the *Construct Adhoc Query* dialog box without making any date of diagnosis changes, click the **Cancel** button at the bottom of the *Date of Diagnosis* dialog box.

Gender

To specify sex codes for your query, click the **Gender** button in the *Construct Adhoc Query* dialog box. The *Gender Selection* dialog appears.

Click (check) the check box next to any category(ies) you wish.

Click **Cancel** to leave the dialog box with no changes made, or **OK** to make what you've chosen part of your query definition[8]. You return to the *Construct Adhoc Query* dialog.

Race

To specify race codes for your query, click the **Race** button on the *Construct Adhoc* Query dialog box, and the *Race Selection* dialog appears. This dialog functions in the same way as the *Hospital Selection* and *Geographic Unit Selection* dialogs.
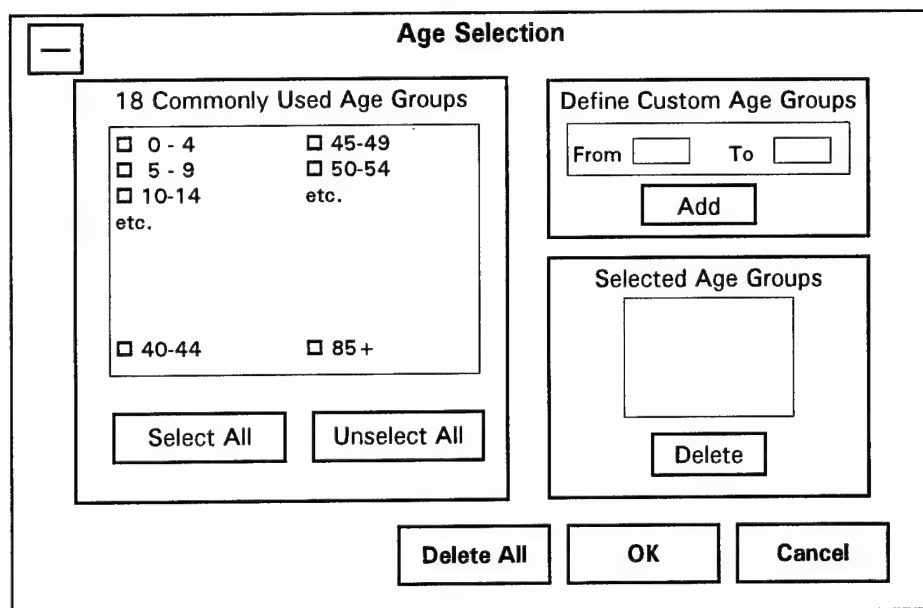
Smoking

To specify patient smoking status codes for your query, click the **Smoking** button on the *Construct Adhoc Query* dialog box, and the *Smoking Selection* dialog appears. This dialog functions in the same way as the *Gender Selection* dialog.

---

[7] NOT the **Delete All** button at the bottom of the dialog!

[8] The only ways to exit this dialog are with **Cancel** or **OK**.

Age

To specify patient age at diagnosis, click the **Age** button in the *Construct Adhoc Query* dialog box. The *Age Selection* dialog appears.

```
┌─────────────────────────────────────────────────────────────────┐
│ ┌──┐                      Age Selection                          │
│ │─ │                                                             │
│ └──┘                                                             │
│   ┌───────────────────────────┐   ┌───────────────────────────┐ │
│   │ 18 Commonly Used Age Groups│   │ Define Custom Age Groups  │ │
│   │  ┌──────────────────────┐ │   │                           │ │
│   │  │ ☐ 0 - 4    ☐ 45-49   │ │   │ From ┌────┐  To ┌────┐   │ │
│   │  │ ☐ 5 - 9    ☐ 50-54   │ │   │      └────┘     └────┘   │ │
│   │  │ ☐ 10-14    etc.      │ │   │         ┌────────┐        │ │
│   │  │ etc.                 │ │   │         │  Add   │        │ │
│   │  │                      │ │   └───────────────────────────┘ │
│   │  │                      │ │   ┌───────────────────────────┐ │
│   │  │                      │ │   │    Selected Age Groups    │ │
│   │  │ ☐ 40-44    ☐ 85+     │ │   │     ┌──────────────┐      │ │
│   │  └──────────────────────┘ │   │     │              │      │ │
│   │  ┌──────────┐ ┌──────────┐│   │     │              │      │ │
│   │  │Select All│ │Unselect  ││   │     └──────────────┘      │ │
│   │  │          │ │  All     ││   │       ┌────────┐          │ │
│   │  └──────────┘ └──────────┘│   │       │ Delete │          │ │
│   └───────────────────────────┘   └───────────────────────────┘ │
│              ┌───────────┐ ┌────────┐ ┌────────┐                 │
│              │Delete All │ │  OK    │ │ Cancel │                 │
│              └───────────┘ └────────┘ └────────┘                 │
└─────────────────────────────────────────────────────────────────┘
```

The order in which you specify ages may be important. If you want the age groups to appear in a certain order in your query results, be sure to specify them here in that order. They will also be listed in the query text box in the order in which you specified them.

To use the 18 standard age groups in your query, click the **Select All** button. To select only certain of these age groups, click on (check) the check boxes of those groups you want, or click **Select All** and then *un*check those groups you *don't* want. Any standard age groups checked will appear in the **Selected Age Groups** box.

To clear *all* the groups you've checked in the **18 Commonly Used Age Groups** area, click the **Unselect All** button.

To specify any age group that is *not* one of the 18 standard, type the range ends in the **From** and **To** boxes of the **Define Custom Age Groups** area and click the **Add** button. The age group you specified will appear in the **Selected Age Groups** box.

To remove any group you've put into the **Selected Age Groups** box, select it and click the **Delete** button underneath the box.[9]

You cannot query on a single age. You cannot have an age in 2 different groups in the same query; that is, you can't group ages 18-65 and 50-65 in the same query.

One of the standard groups is "85 + " (patients aged 85 and older at diagnosis). To specify all ages above (and including) another age, type that beginning age in the **From** box and "200" in the **To** box. That "200" automatically makes CIMS produce a "# + " group. For example, to specify a query for all adults, **Add** the custom group **From 18...To 200**; the age group "18 + " will appear in the **Selected Age Groups** box (and in the query text and query results).

To destroy all the age groups you've specified, click the **Delete All** button and all boxes onscreen will become blank (you can change your mind).

To exit the *Age Selection* dialog box without making any changes, click the **Cancel** button and you will return to the *Construct Adhoc Query* dialog.[10]

---

[9] <u>NOT</u> the **Delete All** button at the bottom of the dialog!

When you have gotten the desired age groups into the **Selected Age Groups** box, click the **OK** button. You will return to the *Construct Adhoc Query* dialog box.

Cancer

To specify primary site, histology and/or behavior codes[11] for a query, click the **Cancer** button in the *Construct Adhoc Query* dialog box. The *Cancer Selection* dialog appears (after some time).

To specify a primary site, you may: select a range of ICD-O-2 codes from the list in the **Site Selection** area and click the **Include** button or **Exclude** button; or double click a range of codes in the list to access its subcodes (a boxed " + " next to an item in the list indicates that it can be expanded into subcodes; a boxed "-" next to an item in the list indicates that it has been broken down in this way; a featureless box next to an item indicates that it can be subdivided no further; double clicking a boxed "-" next to an item will regroup its subcodes); or type the range you want into the **Define Custom Range** area[12] and click **Include**.

To specify your histologies of interest, use the **Histology Selection** area and follow the same procedures as for primary site selection.

Codes that you **Include** will appear in the **Inclusion List** box, and codes that you **Exclude** will appear in the **Exclusion List** box. The codes in the **Exclusion List** box are excluded from the codes in the **Inclusion List** box.[13]

To modify your **Inclusion List** or **Exclusion List**, select the item(s) you want to change in either list and click the **Delete** button next to the list[14]. The code(s) you selected will move back into the original list.

When specifying the behavior code(s)[15] you want (they're in the bottom right of the **Histology Selection** area), be sure to uncheck any behavior codes you wish to exclude from your query definition. Specified behavior codes do *not* appear in the **Inclusion List** or **Exclusion List**, but they are listed in the query text of the *Construct Adhoc Query* dialog box. Note that any checked behavior codes are aggregated, i.e., if you want to see separate observed counts for *in situs* and malignancies for a particular cancer type, you must create 2 different cancer definitions.

When you have gotten the desired codes into the **Inclusion List** ("minus" any codes you've put into the **Exclusion List**) and have the desired behavior codes checked, you <u>must</u> give this cancer definition a name. Type some name (up to any 40 characters) into the **Cancer Name** box, and then click the **Add Cancer** button; the name will then appear in the **Cancer List** drop-down list box and the definition will be retrievable from there.

When you have created at least one cancer definition, you may click the **Group All Others** check box. When the query is run, this feature will calculate the difference between the total observed case count and the total case count for the cancer type(s) you've defined.

To completely remove (destroy) a definition you've named, choose its name from the **Cancer List** box and click the **Delete Cancer** button (you can change your mind).

---

[10] The only ways to exit the *Age Selection* dialog are with **Cancel** or **OK**.

[11] Note that the histology and primary site codes here are not identical to the ICD-O-2 codes which should be here, so we can't really create the cancer definitions we'd like to yet.

[12] Be careful of what you type into the **Define Custom Range** area, because there are few checks on whether what you type here is a non-existent code.

[13] You don't *have to* use the exclusion feature if you don't like it. Use it when it seems convenient. It may sometimes be easier to specify something like "{A-Z} except Q" rather than having to specify "{A-P} and {R-Z}". It's up to the user.

[14] <u>NOT</u> the **Delete All** button at the bottom of the dialog!

[15] Why are codes 4, 5, 7 and 8 there? If these codes ever become defined in the future, CIMS will be ready for them. Our database will only accept codes 0, 1, 2 and 3 anyway (6's and 9's become 3's, and 0's and 1's are only for certain brain/CNS cases; see the new Manual) so just ignore the codes that shouldn't be there.

To clear the screen and start over with a clean slate, click the **Clear Cancer** button. This will not affect any named cancer definitions in the **Cancer List** box.

To modify a cancer definition you've created, click on its name in the **Cancer List** box, change the **Inclusion List, Exclusion List** and/or behavior codes as desired, and click the **Add Cancer** button again to replace the old definition. (You can change your mind and leave the old definition intact.)

To change a cancer definition's *name*, retrieve the definition from the **Cancer List** box, change the name in the **Cancer Name** box to whatever you wish, and click the **Add Cancer** button. Then retrieve the old name from the **Cancer List** and click the **Delete Cancer** button.

To delete *all* the cancer definitions and names in the **Cancer List**, click the **Delete All** button (you can change your mind).

To exit the *Cancer Selection* dialog box and return to the *Construct Adhoc Query* dialog as if you had not entered the *Cancer Selection* dialog, click the **Cancel** button.[16]

When you are satisfied with the definitions you've created, click the **OK** button and you will return to the *Construct Adhoc Query* dialog box.

Industrial Code / Occupation Code

We cannot yet specify values for these data fields.

Saving or Deleting a Query Definition

Saving a Query[17]

When you have specified values for all the data fields you want in your query (at least 2), you may save this query definition by clicking the **Save** (or **Save As**) button at the bottom of the *Construct Adhoc Query* dialog box. Type a name for the query in the box that appears (up to any 70 characters), and then click the **OK** button to save the query definition or the **Cancel** button to return to the *Construct Adhoc Query* dialog. You cannot specify a drive or other location where the query definition is saved (CIMS likes to look after these itself). If you want the query definition to be saved to a floppy or some other specific place, move it with your Windows File Manager.

To save under a new name a previously saved query definition *without replacing the earlier version*, click the **Save As** button; using the **Save** button will automatically replace the old version of the query definition with the current version under the same name.

Deleting a Query

To clear the NEW REGULAR QUERY text box completely so that you can start a new query definition with a clean slate, click the **Clear** button. This does <u>not</u> delete or alter any <u>saved</u> query definitions, but an unsaved query definition will be lost.

To delete (from disk) a saved query definition, click the **Delete** button while that query definition appears in the NEW REGULAR QUERY text box.

---

[16] The only ways to exit this dialog are with **Cancel** or **OK**.

[17] Saving a query definition is optional. You need not save a query definition in order to run that query.

Defining the Query Report

Before you can run a regular query, you <u>must</u> specify how you want the resulting table(s) of data to be laid out onscreen. Click the **Define Report** button at the lower left of the *Construct Adhoc Query* dialog box. The *Report Definition* dialog appears.
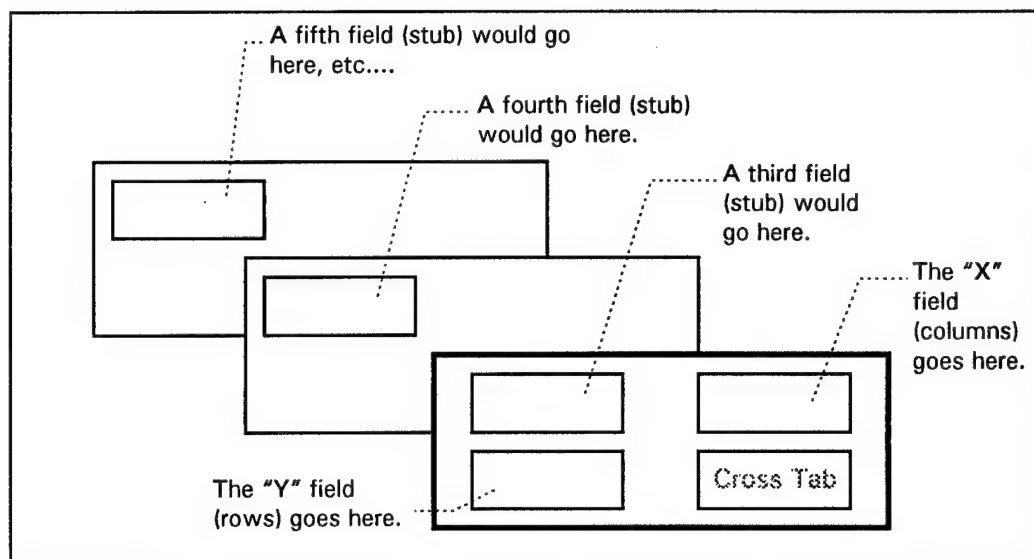
placement of data fields in the *Report Definition* dialog's layout area

Drag the data field names from the **List of Fields** box to the positions where you want them to appear in the layout area on the right. For any report layout, you <u>must</u> drag one data field into the "X" box and another into the "Y" box; stub fields are optional.

You may type a title for any data report you design in the **Report Title** box (with an unlimited number of any characters). The centered lines of text in this box will wrap automatically, and a hard return [Ctrl + Enter] moves the insertion point down to the next line.

Whether or not you have given a title to a report layout, you <u>must</u> click the **Add Report** button to save a layout. The report number (not title) will then appear in the drop-down list in the **List of Reports** box and be retrievable from here.

If you want to delete a report layout you've created, retrieve it by selecting its number from the **List of Reports** box and then click the **Delete Report** button[18] (you can change your mind).

To clear the screen and start over with a clean slate, click the **Clear Report** button. This does not delete or alter any report layouts in the **List of Reports**.

To delete *all* the report definitions you've created for a query, click the **Delete All** button (you can change your mind).

When you've finished laying out your data table(s) for the current query, click the **OK** button. You will return to the *Construct Adhoc Query* dialog box. Your report definitions will be saved and retrievable from the *Report Definition* dialog. When you've retrieved a saved regular query, all the report definitions that existed when you saved it will appear in the numbered **List of Reports** drop-down list box.

To cancel everything you've done in the *Report Definition* dialog box, click the **Cancel** button.

---

[18] <u>NOT</u> the **Delete All** button at the bottom of the dialog!

62

Running a Query

When you have specified values for at least 2 data fields and have created at least one report definition for a query, you can run the query by clicking the **Execute** button in the *Construct Adhoc Query* dialog box. (You'll be told that the query's running; click the **OK** button [Enter] in that little message box.) Then, if you click the **Close** button in the *Construct Adhoc Query* dialog, the query will be shown running in the first available frame in the window full of frames. The percentage that appears in the frame shows how much of the run is completed.

You may halt all running queries and close them by choosing **Query -- Close All Queries** [Alt + Q, C] (you can change your mind).

To temporarily halt a running query, click in its frame and choose the command **Pause Query**. To set a paused query running again from where it left off, click in the query's frame and choose **Resume Query**.

To stop and close a running query, click in its frame and choose **Kill Query** (you <u>cannot</u> change your mind about this).

To change the speed of a running query, click in its frame, choose **Prioritize Query**, and click on a number in the *Query Priority* dialog box (the fastest speed is 8). Then click the **OK** button (to change the speed setting as you've indicated) or **Cancel** (to return to the original speed setting).

To remind yourself of a query definition or get a hard copy of it, click in its frame and choose **View Query Text**. You may click the **Print** button to dump the query definition text to the printer chosen in your Print Manager. <u>You cannot alter the query definition from this dialog box</u>. Use the **Close** button [Esc or Alt + F4] to leave the *Query Text* dialog and return to the window full of frames.

Once a query has reached 100% completion, you may remove it from a frame[19] by clicking in its frame and choosing **Delete Query**. This does *not* remove the query definition from disk if it's been saved.

Viewing a Query's Results

To see the results of a completed query run, click in the query's frame and choose **Reports**. For each report definition you created for this query, the appropriate number of data tables appears in the drop-down list box on the *Query Report* screen. Notice that there is no **Print** button or menu command in this dialog box, so you cannot print the data tables resulting from a regular query directly from CIMS. (You can save the data and print it out from Excel in spreadsheet form).

To view the onscreen tabular results of the run in the form of a graph, click the **Display Graph** button at the bottom of the *Query Report* screen. You may type a title to appear on the graph in the **Graph Title** box (up to any 70 characters). You must select either the rows or columns of the data table to be graphed by clicking either option button in the **Select** area of the *Graph Configuration* dialog box. Then click **OK**, and the *GRAPH* dialog appears. Select a type of graph (the default is a 2-dimensional bar graph). {If you choose a line graph, you may also choose certain statistical features to be indicated on the graph (clicking **Cancel** in this dialog produces a line graph with no such features indicated).} If you want to print a graph, using the **Print** button allows you to control the print setup features. To exit the *GRAPH* dialog, use the **Close** button [Esc or Alt + F4]. You will return to the *Query Report* screen. To view the graph of another table of data that resulted from your query run, choose that table's number from the list in the upper left of the *Query Report* dialog box and then click the **Display Graph** button again.

---

[19] This is optional. A query can remain in its frame for as long as you are in the Adhoc Application. There are 32 frames in the window full of frames that can be filled up with queries.

To save a query's results into a file[20], click the **Save Report** button in the *Query Report* dialog box. A message box will appear asking if you want to save the data with a .dbf format, although you really have no choice in this matter[21]. Choose the **Yes** button, click **OK** in the next dialog box, and you will be reminded that the data in each report defined for this query must be given a different file name. A *Save As* dialog will then appear. Choose a drive/destination and type a file name for the first report, click **OK**, then choose a drive/destination and file name for the second report, etc. If you do <u>not</u> want to save the data from a particular report, **Cancel** out of the *Save As* dialog for that report. When you have saved all the report definitions you want, you will return to the *Query Report* dialog box.

To review or print the text summarizing a query definition, click the **View Query Text** button in the *Query Report* dialog box, and then **Close** [Esc or Alt+F4] your way back out of that dialog.

To sort, annotate or modify a data table in any way, you must do so outside of CIMS. Because you may want to bring a query's data results into Excel, CIMS gives you the option of running Excel directly from the *Query Report* screen. To run the version of Microsoft Excel on your PC, click the **Load Excel** button[22]. If you **Load Excel** after you've saved any query results, Excel will automatically open the first data table you saved. When you exit Excel, you will return to the *Query Report* screen in CIMS.

To exit the *Query Report* screen, use its **Close** button or control-menu box [Esc or Alt+F4].

Retrieving a Saved Query Definition

To retrieve an old query definition that you had saved previously, from the window full of frames, choose **Query -- Retrieve Regular Query** [Alt+Q, R]. The *Retrieve Query* dialog box appears, containing a list of the regular queries in your CIMS "save to" area and the date on which each was last saved.

Double click on one of these query names to open its definition [Tab until the first query listed is highlighted, then ↓ until the one you want to open is highlighted, and then Enter] or click the **Cancel** button [Esc] to return to the window full of frames.

If you open a saved regular query, the *Construct Adhoc Query* dialog box will appear and the query definition will be in the RETRIEVED REGULAR QUERY text box. You may now **Define Report**(s) and/or **Execute** the query, modify the query definition, **Save As** the query with a new name, **Clear** the text box and start defining a new query (this won't affect any saved queries), **Delete** the query definition from disk, **Close** back to the window full of frames, or whatever else you want.

---

[20] This is optional. Saving the data saves you from having to re-run the same query later to obtain the same results.

[21] If you choose the **No** button, the query results will not be saved at all. If your data must end up in a different file format, save it here as a .dbf file and transform it to the format you want outside of CIMS (maybe with Howard's help).

[22] If nothing happens after you click **Load Excel**, it probably means that your PC lacks the memory to run CIMS and Excel simultaneously. If this happens, make sure you save any report results you want as .dbf files; you can later <u>close CIMS</u> and then open the .dbf files in Excel.

STATISTICAL QUERIES

Statistical queries produce SIRs, age-specific incidence rates and age-adjusted incidence rates (all in one query).

Defining a New Statistical Query

To define a new statistical query, from the window full of frames, choose **Query -- New Statistical Query** [Alt + Q, E]. The *Construct Adhoc Query* dialog box appears with only 3 command buttons available at the top -- **Geographic Unit, Diagnosis Date** and **Cancer**.

Specifying Data Field Values

Defining a statistical query is initially simpler than defining a regular query because you can only specify values for 3 data fields (geographic unit, date of diagnosis and cancer type), but you <u>must</u> specify values for <u>all</u> 3 fields[23]. CIMS automatically specifies male, female and the 18 standard patient age groups. You cannot change the sex selection, but Total (= Male + Female) data also appears in the query results automatically (the transsexual, hermaphrodite and unknown categories are excluded). You can change the patient age groupings <u>after</u> the query has been run with the 18 standard groups.

Geographic Unit

The only differences between specifying geographic units here and in a regular query is that statewide, census tract and ZIP Code data are not available for a statistical query, and units cannot be grouped. Choose individual (**Non-Group Items**) towns, counties or CHNAs as you would for a regular query.

Diagnosis Date

In a statistical query, you can choose only *individual entire* years of data initially. (You can aggregate entire years of data <u>after</u> the query has been run.) The **Select by Dates** area of the dialog is therefore unavailable during a statistical query. Otherwise, the dialog functions as it does for a regular query.

Cancer

Specify this in the same way as for a regular query.

Patient Age

The default is to the 18 standard age groups. You can change these age groupings <u>after</u> the query has run (when you are viewing the query results).

Saving or Deleting a Query Definition

Either function is performed from the *Construct Adhoc Query* dialog box (as for a regular query).

Defining Reports for a Statistical Query

You don't design the layout of your query results, so the **Report Definition** button in the *Construct Adhoc Query* dialog box is not available. CIMS displays all statistical query results in automatic layouts.

---

[23] You need only specify values for **2** data fields in a regular query.

For SIRs, the table layout consists of these columns:

<u>cancer type</u>   <u>sex</u>   <u>observed count</u>   <u>expected count</u>   <u>SIR</u>   <u>confidence interval</u>

For age-specific rates, the columns are:

<u>cancer type</u>   <u>age group</u>   <u>sex</u>   <u>observed count</u>   <u>rate</u>   <u>confidence interval</u>

For age-adjusted rates, the columns are:

<u>cancer type</u>   <u>sex</u>   <u>observed count</u>   <u>rate</u>   <u>confidence interval</u>

Running a Query

The only way this is different than running a regular query is that 2 frames in the window full of frames are needed to display the status of a statistical query. One of these frames represents the query you've defined, and the other represents a corresponding statewide query. You can control running and completed statistical queries as you can regular queries. The "priority" (speed) you give either frame of a running statistical query applies to the other frame as well.

When both query frames reach 100% completion, click on either frame and choose **Reports**. A dialog box entitled *Statistical SubSystem* appears (after some time)[24]. It has these menus and commands:

File	**Setup Statistic Controls**	**View**	**Help**
Open Ctrl+O	Confidence Limits	✓Toolbar	About Statapp...
Save Ctrl+S	Age Group	✓Status Bar	
Save As	Data Years		
-----------------------	Pop. Mid-Year Selection		
Print Ctrl+P			
Print Preview			
Print Setup...			

Exit [Alt+F4]			

Beneath the menu bar may appear a toolbar containing buttons that are not of much use and a drop-down list of the types of statistics available (SIR is the default).

Below this is the **Tabulation of result from statistical computation** area, containing drop-down lists for the year(s) of diagnosis and geographic unit(s) you specified. You can change the column widths by dragging them so that you can see more data onscreen without having to scroll.

Changing Defaults

While viewing the results of a statistical query, you can make certain changes to the statistics displayed. You can view SIRs, age-specific rates or age-adjusted rates (one at a time), select confidence limits, change age groupings, group diagnosis years, and select the populations used (denominators).

Choosing Statistics

The system by default displays SIRs when you first view a statistical query's results. You can change the statistic being displayed to age-specific rates or age-adjusted rates by pulling down the list in the box that initially says **Standardized Incidence R...** and selecting one of the other statistics.

---

[24] If you get a message about graphics.exe not being available vis the DOS path, your PC probably lacks sufficient memory for you to view these query results. Ask for help in freeing up memory, or run the query at someone else's workstation.

Selecting Confidence Limits

The default setting is 95%. You may change the setting to 90% or 99% by choosing **Setup Statistic Controls -- Confidence Limits** [Alt + S, L] from the *Statistical SubSystem* dialog box and clicking either of these option buttons. Then click **Close** [Enter] to change the setting to what you've selected, or **Cancel** [Esc or Alt + F4] to leave the setting as it was. The data will be recalculated and you will return to the *Statistical SubSystem* dialog box.

Regrouping Ages

The default is the 18 standard age groups. To change this, choose **Setup Statistic Controls -- Age Group** [Alt + S, A] from the *Statistical SubSystem* dialog box, and the *Edit Age Group* dialog appears.

The current age groups can be viewed in the drop-down list box in the **Selected Age Group** area. To modify these groupings, click an option button[25] under **From Age**, an option button under **To Age**, and then click **Add Group**. The new group will move into the **Selected Age Group** list. Repeat these steps until you've grouped all the ages as you wish. Each age must be included in one and only one group. You may find that you cannot group together exactly the ages you want because CIMS limits you to the ages that appear in front of the option buttons.

To change what you've grouped, select a group from the list in the **Selected Age Group** area and click **Remove Group**. Those ages will return to the top of the screen.

When you've gotten all the ages grouped as you want, click **Close**. (Although there is no **Cancel** button, you can "cancel out" of the dialog using [Esc or Alt + F4] to leave the original age groups intact.) The data will be recalculated and you will return to the *Statistical SubSystem* dialog box.

Aggregating Years of Diagnosis

The default is to the individual entire years you chose when defining the query. Each of these years begins as a "group" unto itself[26]. To modify this, choose **Setup Statistic Controls -- Data Years** [Alt + S, Y] from the *Statistical SubSystem* dialog box, and the *Edit Group of Data Years* dialog appears.

The years specified can be viewed in the drop-down list box in the **Selected Year Groups** area. Select a year from this drop-down list and click **Delete Group**. Each year you selected is "ungrouped" and moves into the **Single Year List** box.

To aggregate a series of consecutive[27] years, select the first and last years in the series in the **Single Year List** box and click the **Add→** button; they will move into the **Selected Years** box. (To move a year back, select it in the **Selected Years** box and click **←Remove**.) When you've gotten the years you want to aggregate into the **Selected Years** box, click **Make Group**. The aggregated years will move down to where they began -- the **Selected Year Groups** drop-down list box.

Repeat these steps until you've created all the groups you want. When done, click **Close**. (To "cancel out" of the dialog without making any changes, use [Esc or Alt + F4] because there is no **Cancel** button.)

---

[25] Note that the option button for an age <u>follows</u> (is to the right of) that age.

[26] For example, if you chose 1982, 1983 and 1984 when defining the query, you begin with 3 groups consisting of those individual years.

[27] You can only aggregate consecutive years of data, i.e., if you specified only 1984 and 1986 data when defining the query, data for those years must remain separate.

Specifying the Population Year

Choose **Setup Statistic Controls -- Pop. Mid-Year Selection** to choose your rate denominators. The *Select Mid-Year For Statistical Computation* dialog box appears.

Printing Statistics Tables

You can<u>not</u> print the contents of an onscreen data table from within CIMS! Right now you can only view query results onscreen. At some point we should be able to save these in .dbf format and print and manipulate them in Excel.

Saving Statistics Tables

You can<u>not</u> save a statistical query's data table by choosing **File -- Save** or **File -- Save As** from the *Statistical SubSystem* dialog box menus! Right now we can only view query results onscreen. At some point we should be able to save these in .dbf format.

The Graph & Map Screens

When viewing data on the "spreadsheet" screen, there are 2 other screens to its left that can be used to view the data differently. If you move the cursor along the extreme left of the dialog (where there are 2 tall bars) it will change shape to ←‖→. If you drag one of these cursors toward the right, a Mass. map will be revealed. You can choose the type of statistic you want to view from the drop-down list box under the menu bar, or change the specific data {diagnosis year(s), cancer type, sex, age group (for rates)} being viewed on the map by pulling down lists in the **Map Control Panel**. You can also change the **Setup Statistic Controls** (Confidence Limits, Age Group, Data Years, Mid-Year Selection) here, and any changes made on this screen will also affect the spreadsheet and graph views of the data. Clicking in a town on the map will reveal its name.

Clicking the **Cancer Statistics** button reveals a *Statistical Legend* box with the data being viewed and (for rates) the corresponding population. The arrow buttons next to the **Close** button move you through your geographic units. Click the **Close** button in this little box to return to the map or graph screen.

Pulling the other ←‖→ cursor on the left of the screen toward the right will uncover a graph window. You can again change what data from the query you're viewing and/or change the **Setup Statistic Controls**. There is also a **Cancer Statistics** button here (like the one on the Map screen).

The graph screen is on the far left, the "spreadsheet" screen is on the far right, and the map screen is in the middle of the arrangement.

When you've finished viewing the results of your query, you can exit the *Statistical SubSystem* dialog box and return to the window full of frames by using the control-menu box [Alt+F4] or by choosing **File -- Exit** [Alt+F, X].

Retrieving a Saved Query Definition

This works the same way as retrieving a saved regular query. From the window full of frames, choose **Query -- Retrieve Statistical Query** [Alt+Q, S] to get started.

LEAVING THE *CONSTRUCT ADHOC QUERY* DIALOG

When you want to leave the *Construct Adhoc Query* dialog box, you must use the **Close** button. (Its control-menu box [Alt+F4] will <u>not</u> close it.) You will return to the window full of frames.

CLOSING THE ADHOC

To exit the Adhoc Application, go to the window full of frames and choose **Query** -- E**x**it [Alt+F4] or use the control-menu box of the window full of frames.

This is a step-by-step exercise for creating a regular query definition with these specifications:

breast cancer
in Boston vs. Worcester + Springfield combined
for females under 35 years of age, 35-39, 40-44, 45-49 and over 49
for time periods June 15 - December 22, 1982; 1986; and 1988-1990

(You can create this definition in other ways. These steps just show how I might go about it.)

Run the Adhoc Application.
Choose **Query -- New Regular Query**.

Click **Geographic Unit**.
Click **City/Town**.
Click **Abington**.
Type [B].
PageDown to **Boston**.
Drag **Boston** into **Non-Group Items** box.
[T]
PageUp to **Springfield**.
Drag **Springfield** into **Group Items** box.
[Y]
Drag **Worcester** into **Group Items** box.
Type into **Group Name** box "Worc & Spfld".
Click **Add Group**.
OK

Click **Diagnosis Date**.
Click the first ("From") 15 in **Select by Dates**.
Click in the text box where 1996 appears and change the last two digits to "82".
Click on June 15.
OK
Click the "To" 15.
Click where 1996 appears and change to1982.
Click December 22.
OK
Click **Add**.
Click **1986** in **Select by Years**.
Click **Select**.
Type 01/01/1988 into the "From" box.
Type 12/31/1990 into the "To" box.
Click **Add**.
OK

Click **Gender**.
Click **Female**.
OK

Click **Age**.
Type into **Define Custom Age Groups:** "0" in the From box, "34" in the To box.
Click **Add**.
Click the **35-39** option box.
Click the **40-44** option box.
Click the **45-49** option box.
Type into **Define Custom Age Groups:** "50" in the From box, "200" in the To box.
Click **Add**.
OK

Click **Cancer.**
Scroll down the list of primary site codes to **C50.**
Click **C50.**
Click **Include.**
Type into **Define Custom Range:** "8000" in the From box, "9999" in the To box.
Click **Include.**
Type into **Define Custom Range:** "9590" in the From box, "9980" in the To box.
Click **Exclude.**
"Unclick" all behavior codes except **3.**
Type into **Cancer Name** box "Adhoc example (BR)".
Click **Add Cancer.**
OK

Click **Define Report.** Let's create two reports that present the data in slightly different ways.
Drag **City/Town** to Y position.
Drag **Age** to X position.
Stub on **Diagnosis Date.**
Type **Report Title** "to produce 3 data tables".
Click **Add Report.**
Drag **Age** to Y position.
Drag **Diagnosis Date** to X position.
Stub on **City/Town.**
Type **Report Title** "2 tables for this report".
Click **Add Report.**
OK

Click **Save.**
Type file name "Mary's regular example".
OK

Click **Execute.**
OK

Click **Close.**

While the regular query runs, try defining this statistical query:

SIRs for Cape Cod prostate cases for 1982-1986 and 1987-1992, and a 1990 age-specific incidence rate for Cape males over 64

Choose **Query -- New Statistical Query.**

Click **Geographic Unit.**
Click **County.**
Drag **Barnstable** into **Non-Group Items** box.
OK

Click **Diagnosis Date.**
Click **1982.**
Shift-click **1992.**
Click **Select.**
OK

Click **Cancer.**
Scroll down primary site code list.
Double click **C60-C63.**
Double click **C61.**
Click **C61.9.**
Click **Include.**
Type into **Define Custom Range:** From "8000", To "9999".
Click **Include.**
Type into **Define Custom Range:** From "9590" To "9980".
Click **Exclude.**
"Unclick" all behavior codes except **3.**
Type **Cancer Name** "Cape Prost".
Click **Add Cancer.**
OK

Click **Save.**
Type file name "Mary's stat example".
OK
Click **Execute.**
OK
Click **Close.**

When the regular query has finished running, click in its frame and choose **Reports**.
Look at the layout of the 3 tables for Report 1.
Click the option button for **Report 2** and look at its 2 tables.
Click **Close**.

When the statistical query has finished running, click in its frame and choose **Reports**.

Choose **Setup Statistic Controls -- Data Years**.
Click **Delete Group** until all years have moved from **Selected Year Groups** to **Single Year List**.
In **Single Year List**, select **1982-1982**, then select **1986-1986**.
Click **Add→**.
Click **Make Group**.
In **Single Year List**, select **1987-1987**, then select **1992-1992**.
Click **Add→**.
Click **Make Group**.
Click **Close**.
Look at the 1982-1986 SIRs....
Pull down the **Date of Diagnosis** list and choose 1987-1992.
Look at the 1987-1992 SIRs....

When you've finished looking at the SIRs, pull down the list that says "Standardized Incidence R...".
Choose **Age-Specific Rate**.

Choose **Setup Statistic Controls -- Age Group**.
In the **From Age** area, click the option button following "65".
In the **To Age** area, click the last option button (for "85 + ").
Click **Add Group**.
Click **Close**.

Choose **Setup Statistic Controls -- Data Years**.
Pull down the **Selected Year Groups** list.
Choose 1987-1992.
Click **Delete Group**.
In the **Single Year List** box, choose 1990-1990.
Click **Add→**.
Click **Make Group**.
In the **Single Year List** box, choose 1987-1987, then choose 1989-1989.
Click **Add→**.
Click **Make Group**.
In the **Single Year List** box, choose 1991-1991, then choose 1992-1992.
Click **Add**.
Click **Make Group**.
Click **Close**.

Pull down the **Date of Diagnosis** list.
Choose 01/01/1990-12/31/1990.
Scroll down to the 65-85 + male data and look at the rate.

Use the control-menu box to close out to the window full of frames.

QUALITY CONTROL CHECK ON GEOCODED DATA

The following notes refer to the quality control check of the geoding done by an outside contractor. A sample of 473 cases were selected from towns in Massachusetts begining with "D" (Dalton - Dracut) and MAPINFO was utilized to manually examine these cases.

	# INCORRECT		- # UNEXPECTED	= ERROR RATE (POBox, institution, not enough info)		MAPINFO RATE	
DALTON	(7/21)	33.3%	(3/21)	(4/21)	19.0%	3/21	14.3%
DANVERS	(27/72)	37.5%	(14/72)	(13/72)	18.1%	3/72	4.2%
DARTM.	(17/75)	22.7%	(5/75)	(12/75)	16.0%	4/75	5.3%
						*[11/75	14.7%]
DEDHAM	(13/98)	13.3%	(3/98)	(10/98)	10.2%	5/98	5.1%
DEERFIELD	(11/18)	61.0%	(6/18)	(5/18)	27.8%	0/98	0%
DENNIS	(48/90)	53.3%	(20/90)	(28/90)	31.1%	2/90	2.2%
						*[15/90	16.7%]
DIGHTON	(4/17)	23.5%	(0/17)	(4/17)	23.5%	2/17	11.8%
DOUGLAS	(5/11)	45.5%	(1/11)	(4/11)	36.4%	2/11	18.2%
DOVER	(2/16)	12.5%	(0/16)	(2/16)	12.5%	1/16	6.3%
DRACUT	(8/55)	14.5%	(0/55)	(8/55)	14.5%	3/55	5.5%
Totals:	(138/473)	31.6%		(90/473)	19.0%	25/473	5.3%
						45/473	9.5%

*these are cases in which I was unable to identify the census tract using MAPINFO
because the street crosses more than one census tract.

74

PERCENTAGE OF CASES MISSING THE
CENSUS TRACT

DALTON	2/21	9.5%
DANVERS	15/72	20.8%
DARTM.	14/75	18.7%
DEDHAM	6/98	6.1%
DEERFIELD	7/18	38.9%
DENNIS	33/90	36.7%
DIGHTON	1/17	5.9%
DOUGLAS	4/11	36.4%
DOVER	2/16	12.5%
DRACUT	5/55	9.1%
Totals:	89/473	**18.8%**

The most common errors are:

    *   No census tract & wrong lat. and long.   (84)

    *   Wrong census tract & wrong lat. and long   (19)

    *   Wrong zipcode   (9)

    *   Wrong lat. and long. (5)

    *   Changed the name of the street (5) (#'s: 79,198, 207,210,327)

    *   Wrong census track, lat.,long &zip (2)

    *   Wrong lat., long., & zip (1)

***   given no census tract, the lat. & long. are always wrong   ***
In this situation,we are usually given the center of the
MCD or in some towns the center of the census tract.

***   given no house number with the street name the lat. & long. are ***
sometimes wrong
(WRONG: #12, 138,281,291,303,348,369,395,397,400,402,
    OK: #121,227,320,394)
In this situation, we are usually given the center of the MCD or in some towns
the center of the census tract (Dartmouth and Dennis)

***   given a PO Box, no census tract is reported and the lat. & long. are wrong ***
In this situation,we are usually given the center of the MCD or in some towns
the center of the census tract Dartmouth and Dennis).

***   given an institution, no census tract is reported and the lat. & long. are   ***
wrong, or census tract is reported but the lat. & long. are wrong.
In this situation,we are usually given the center of the MCD or in some towns
the center of the census tract (Dartmouth and Dennis).

	CENTER OF MCD		CENTER OF CT	
	LAT.	LONG.	LAT.	LONG.
DALTON (1 CT)	42.47670	73.15530		
DANVERS (5CT)	42.57620	70.95580		
DARTMOUTH (3CT)			41.676645	71.017949
			41.609735	70.985770
			41.556240	70.986945
·DEDHAM (5CT)	42.24460	71.18160		
DEERFIELD (1CT)	42.521545	72.615799		
	42.52460	72.61590		
DENNIS (5CT)			41.732349	70.205074
			41.730570	70.152979
			41.693620	70.156734
			41.667340	70.171139
			41.668410	70.137865
DIGHTON (1CT)	41.836050	71.156405		
	41.83490	71.11930		
DOUGLAS (1CT)	42.054304	71.755160		
	42.05830	71.75530		
DOVER (CT)	42.241576	71.287851		
DRACUT (3CT)	42.692537	71.309050		
	42.69270	71.30760		

Street Names:

* Given: #50 Mass Ave------Not able to determine that this was Massachusetts Ave

* Given: #79 Pines St--------Not able to determine that this was Pine

* Given: #467 Lake View Ave----Not able to determine that this was Lakeview Ave

****

Incorrectly changed the name of:

*	#198 Pleasant ------- to: Boathouse**	(Deedham)
*	#207 Chauncy -------to: Waterview Pl**	"
*	#210 Harding --------to: Cleavland**	"
*	# 327 Pine--------------to: Pond**	(Dennis)

**(these are non existent streets in the respective towns)

****

The following names were changed and are probably correct:

* #58   Proter--------------to: Porter
* # 80 Clifton--------------to: Clinton
* #94   Mylesstandis--------to: Standish
* #96   Strathmor-------------to: Strathmore
* #100 Chace --------------to: Chase
* # 226 Barraus------------to: Barrows
* #460 Turgan-------------to: Turgeon

****

APPENDIX D. PROJECT PUBLICATIONS AND MEETING ABSTRACTS

Sheehan TJ, Gershman ST, MacDougall L, Danley R, Rhoden DH, and Sorensen AM. Using GIS with Cancer Registry and Census Data as a Tool to Monitor Breast Cancer Screening. Poster Session at the Society for Epidemiologic Research Annual Meeting, June 1996.

Sheehan TJ, Gershman ST, MacDougall LA, Danley RA, and Sorensen AM. GIS and Breast Cancer Screening: Integrating Cancer Registry, Census, and Mammography Site Data to Monitor Breast Cancer Control. Presented at the North American Association of Central Cancer Registries Annual Meeting, April 1997.

Sheehan TJ, Gershman ST, MacDougall L, Danley R, Mroszczyk M, and Sorensen AM. GIS and Breast Cancer Screening: Integrating Cancer Registry, Census, and Mammography Site Data to Monitor Breast Cancer Control. Presented at the Joint Meeting of the Public Health Conference on Records and Statistics and the Data Users Conference, July 1997.

Sheehan TJ, Gershman ST, MacDougall L, Danley R, Mroszczyk M, and Sorensen AM. GIS: A Tool to Monitor Cancer Control. Poster session at the Centers for Disease Control and Prevention Cancer Conference: Integrating Public Health Programs for Cancer Control, September 1997.

Sheehan TJ, Gershman ST, MacDougall LA, Danley R, Mroszczyk M, and Sorensen AM. Using Computer Maps to Assess Mammography Screening. To be presented at the Department of Defense Era of Hope Conference, Washington DC, November 1997.

APPENDIX E. PERSONNEL RECEIVING PAY FROM THIS EFFORT

Joan MacDonald, MS

Cuong Nguyen

T. Joseph Sheehan, PhD

Ann Marie Sorensen, MS (cand.)

Hung Tran

Amy Wang, MS

Jim Zazzali, MPH